5-24-2019

Probabilistic and Statistical Prediction Models for Alzheimer's Disease and Statistical Analysis of Global Warming

Maryam Ibrahim Habadi
University of South Florida, habadi_memo@hotmail.com

Probabilistic and Statistical Prediction Models for Alzheimer's Disease and Statistical Analysis of Global Warming

by

Maryam Ibrahim Habadi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a concentration in Statistics
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris P. Tsokos, Ph.D.
Getachew A. Dagne, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Lu Lu, Ph.D.

Date of Approval:
April 23,2019

Keywords: Parametric Analysis, Times Series, Copula, Carbon Dioxide, Beta-Amyloid, Tau Protein

Dedication

I dedicate my dissertation to my loving husband Hani Nader for his endless support throughout the process. To my loving parents Ibrahim Habadi and Badriah Hussain, who sacrificed their valuable time to enlight my future. To my gorgeous daughter Amna and adorable son Hamza, and my wonderful sisters and brothers.

Acknowledgments

I want to express my sincere gratitude to my advisor, Prof. Chris P. Tsokos whom I have had the pleasure to work with for his support, motivation and professional guidance for my Ph.D. research and life in general. He has shown me, by his example, what a good advisor should be.

My sincere thank go to Dr. Getachew Dagne, Dr. Kandethody Ramachandran, and Dr. Lu Lu, for serving in my Ph.D. committee and for their advice and support. Also, many thanks to my friend Mohamed Abu Sheha for his fruitful discussions.

Finally, I would like to give my great thank to my parents and my beloved husband, Hani Nader. Without their endless love, support and encouragement to do my best, I could not have achieved my goal. Thank you all.

Table of Contents

List of Tables

List of Figures

Abstract

The importance and applicability of data-driven statistical models have increased significantly. This current study, we have utilized analytical techniques in interdisciplinary research, including environmental and health.

Environmentally, global warming is considered one of the critical issues facing our planet. It is the increase in average global temperatures caused mostly by increases in Carbon Dioxide CO_2. The excessive rise of carbon dioxide from the average level as the side effect of the industrial revolution has a significant impact on blocking the heat and increase the temperature within the Earth's atmosphere. Based on the record of total CO_2 emissions from fossil fuel burning and cement production in 2014, Saudi Arabia ranked as the 8th largest carbon dioxide emitter among all the countries in the world and some of the Middle Eastern countries are in the top 50.

In the first part of the study, we have developed a data-driven nonlinear statistical model to identify the significant types of fossil fuel (gas fuel, liquid fuel, and solid fuel), cement manufacture, and gas flaring and their possible interactions and have ranked them based on their percentage of contribution to the atmospheric CO_2 concentrations in the Middle East. Then, we compared the results to the findings with those of the United States, the European Union, and South Korea.

Second, the multiplicative seasonal autoregressive integrated moving average (seasonal ARIMA) model is used to develop statistical time series forecasting models to predict carbon dioxide in the atmosphere in the Middle East and atmospheric temperature in Saudi Arabia. Thus, the resulting statistical predictive model is useful in forecasting and monitoring the future level of carbon dioxide emission and extracting meaningful statistics and characteristics about the emission of carbon dioxide in the Middle East.

In health science, Alzheimer's disease is one of the most critical diseases our planet is facing since it is a rapidly increasing disease as the population ages, and the diagnosis of the disease is still poorly understood. Thus, the need for biomarkers for reliable diagnosis is tremendous to help in finding treatment to this severe disease. Hence, the main aim of this study is to utilize information from baseline measurements to develop a statistical prediction model using multiple logistic regression to identify patients with Alzheimer's disease from cognitively normal individuals. Our optimal predictive model includes five risk factors and two interaction terms and has been evaluated using classification accuracy, sensitivity, specificity values, and area under the curve.

Finally, as researchers and scientists suggested that the abnormal level of beta-amyloid and phosphorylated tau ($P\tau$) proteins as one of the possible causes of Alzheimer's, we performed parametric statistical analysis to the beta-amyloid and the $P\tau$ proteins levels of Alzheimer's patients to understand their probabilistic behavior independently. This study involves the identification of the probability distribution function that characterizes the behavior of the subject variables of interest. Having identified such a probability function, we can obtain useful information concerning the two subject entities, such as the expected numerical value and confidence level of the beta-amyloid and P-tau proteins. The second main aim of this study is to explore their probabilistic behavior as correlated variables by establishing their bivariate probability distribution function. A copula method is proposed to model the joint probability density function of both proteins with the given marginals and correlation coefficient. Usually, researchers working on Alzheimer's data characterize the probability distribution function (pdf) of beta-amyloid and P-tau protein levels as the popular Gaussian pdf. The required symmetry of the data is not correct in the subject area, and the results will be misleading. Thus, the best distributions that fit the levels of beta-amyloid and P-tau proteins are the three parameters log-logistic probability distribution and the three-parameter Weibull probability distribution, respectively.

Chapter One

Introduction

In this chapter, we briefly present our research goals of the dissertation in the analytical development and applications environmental and health sciences.

1.1 Carbon Dioxide and Global Warming

What is the relationship between carbon dioxide and global warming? The gases in Earth's atmosphere include 78% nitrogen, 21% oxygen, 0.93% argon and a minimal amount about 0.04% of carbon dioxide is present in the atmosphere. Even though carbon dioxide present in a small percentage, it has a significant impact on sustainable life on the planet. Carbon dioxide (CO_2) plays a crucial role in trapping heat in the atmosphere and keeping our world from freezing. However, the way that humans live, using fossil fuels and other practices that release CO_2 into the air, contributes to the amount of atmospheric carbon dioxide. As carbon dioxide concentrations in Earth's atmosphere continue to increase, meaning add to the amount of heat trapped in the atmosphere, which raises the temperature of the planet "The Intergovernmental Panel on Climate Change has fully documented the fact that industrial activity is responsible for the rapidly increasing levels of atmospheric carbon dioxide and other greenhouse gases. It is not surprising then that global warming can be linked directly to the observed increase in atmospheric carbon dioxide and to human industrial activity in general."(Lacis, 2010). With respect to that, global warming is a function of two main factors in the atmosphere, carbon dioxide, and atmospheric temperature. Many scientists around the world consider global warming as a series problem affecting our planet, and we have to understand its causes to be able to do something to save our beautiful earth.

1.1.1 Statistical Analysis and Modeling of the Atmospheric Carbon Dioxide in the Middle East

Saudi Arabia has been ranked as the eighth largest carbon dioxide emitter among all the countries in the world in addition to some of the Middle Eastern countries are in the top 50, based on the record of total CO_2 emissions in thousand metric tons from fossil fuel burning and cement production in 2014. The objective of our study in Chapter 2 is to develop a data-driven nonlinear statistical model to identify the actual significant attributable variables and their interactions terms that produce carbon dioxide as it is the critical element of global warming.

In this study, we consider fossil fuel burning (gas fuel (Ga), liquid fuel (Li), and solid fuel (So)), cement manufacture (Ce), and gas flaring (Gl) with all possible interactions as risk factors that may contribute to the atmospheric CO_2 concentrations in the Middle East. The different types of fossil fuels and their interactions have been identified and ranked based on their percentage of contribution to CO_2 in the atmosphere. We compared the results of our model with the finding of the United States, the European Union, and South Korea. The developing model and the comparison are useful in structuring regional strategic policies and plans, but not global, to maintain an optimal level of CO_2 in the atmosphere.

In the process of the statistical modeling of the carbon dioxide in the Middle East, we used the coefficient of determination (R^2) and adjusted R squared (R^2_{adj}) criteria to select and evaluate the proposed model. Also, we performed a residual analysis that calculated the actual value of CO_2 in the atmosphere (response) minus the estimated value of CO_2 in the atmosphere using the proposed statistical model, and it attests the quality of the developed statistical model. Additionally, we rank the contribution of risk factors by assessing the relative significance of the risk factors in determining the response variable.

1.1.2 Statistical Forecasting Model of the Atmospheric Carbon Dioxide in The Middle East

Time series is an essential and powerful statistical procedure used to forecast the future vision of

the phenomenon of interest. The chapter aims to develop a statistical time series forecasting models to predict carbon dioxide in the atmosphere in the Middle East and atmospheric temperature in Saudi Arabia. It is known that the excessive rise of carbon dioxide from the average level as the side effect of the industrial revolution has the significant impact in blocking the heat and increase the temperature within the Earth's atmosphere. Thus, the resulting statistical predictive model is useful in forecasting and monitoring the future level of carbon dioxide emission and extract meaningful statistics and characteristics about the emission of carbon dioxide in the Middle East. Also, it assists in developing a strategic policy to maintain the maximum allowable production of carbon dioxide in the Middle East.

In this study, we used monthly data of atmospheric carbon dioxide level measured in part per million from 1996 to 2015 and average monthly temperature measured in Celsius of Saudi Arabia. In developing our statistical predictive models, we used the multiplicative seasonal autoregressive integrated moving average (seasonal ARIMA), that was first introduced by Box and Jenkins (1976) and become the most popular methods for modeling non-stationary time series data. In the case where seasonal components are included in this model, then the model is called seasonal ARIMA (SARIMA). The following steps described the methodology of building the SARIMA model based on the Box-Jenkins procedure,

1. *Model Identification*: obtain data stationarity by differencing and transform data to stabilize variance. Then identify the orders of autoregressive and moving average by plotting and computing the autocorrelation function (ACF) and the partial autocorrelation function (PACF).
2. *Model Estimation*: After getting the stationary of the data, we estimate the parameters using maximum likelihood estimation to minimize the mean square error function.
3. *Model Validation*: include a residual analysis that shows a randomly distributed error with constant mean and variance by checking errors autocorrelation and partial autocorrelation functions. If the residual satisfies these assumptions, then the statistical forecasting model is a good model.
4. *Forecasting*: After validating the model, the best model is used for forecasting future values of the phenomenon of interest.

We will discuss those steps practically on our data.

1.2 Alzheimer's Disease (AD)

Alzheimer's disease is not a normal part of aging and is the most common form of dementia that causes problems with memory, thinking, and behavior. It is the 6th leading cause of death in the United States and the only one that cannot be prevented, treated or even slowed. According to Alzheimer's Association, every 66 seconds someone develops the disease in the U.S and an estimated 5.5 million American are living with AD. Almost two third of Americans with Alzheimer's disease are women, and older African American get the disease twice as older white.

When we confront any problem, we first need to find what causes the problem and then to try to better understand statistically its behavior. It is still unknown what are the risk factors that significantly contribute to Alzheimer's disease, but it is known that Alzheimer's is caused by physical changes, the death of the nerve cells, in the brain. During the first stage of Alzheimer's disease, people seem free of any symptoms, but the toxic changes may occur years or even a decade before we realize the disease presence. Once the healthy nerve cells (neurons) die and lose connections with other neurons, memory loss and other problems occur. As more brain cells die, this leads to significant shirk of the brain size [1]. Figure 1.1 shows the significant differences between the brain size of Alzheimer's patient and a healthy individual.

The autopsies of the brain affected by Alzheimer's disease always show tiny inclusions of the nerve tissue called plaques and tangles. Plaques are found between the dying cells in the brain from the buildup of a protein called amyloid beta and tangles are twisted fibers within the dying cells from another protein called tau. Amyloid and tau proteins are normally fragmenting that the body produces, but in Alzheimer's the proteins are abnormal. Scientists believe plaques and tangles may not be the only factors but considered the main features involved in Alzheimer's disease.

Alzheimer's disease is diagnosed through clinical examination, behavioral assessment to measure the severity of the disease. Besides, advanced brain imaging that allows seeing plaques and tangles in a living brain, blood and fluid biomarkers requires a collection of cerebrospinal fluid that surrounds the brain and extends into the spinal cord. This analysis provides insight into how the disease progresses.

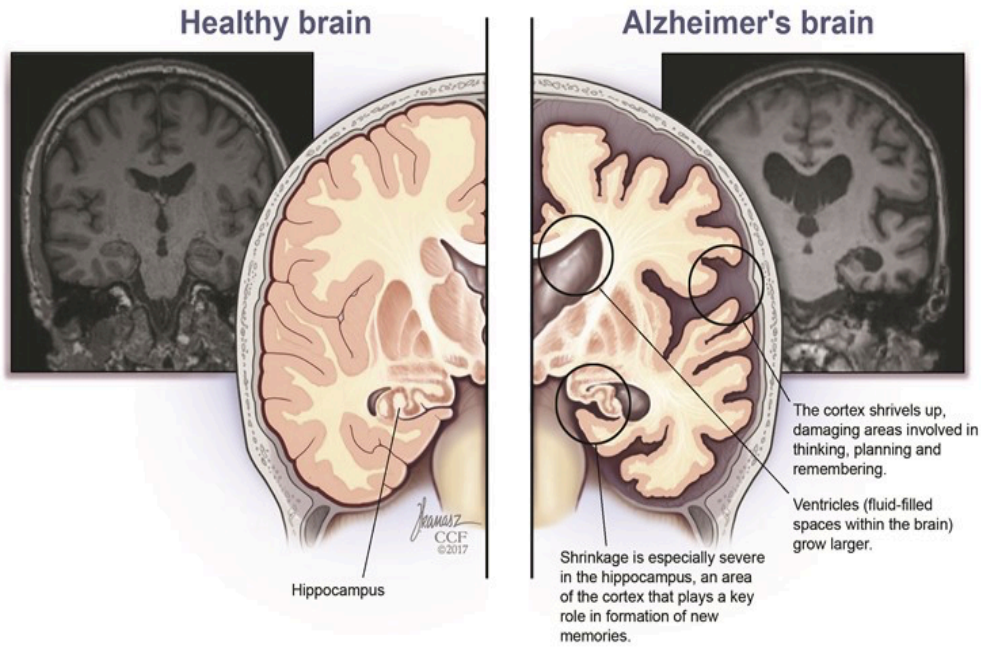


Figure 1.1 Illustration between a brain affected by Alzheimer's disease and a healthy brain

*Source: Keep Memory alive: https://www.keepmemoryalive.org/brain-science/alzheimers-brain

1.2.1 Alzheimer's Disease: The Relative Importance Diagnostic

In the United States, several leading causes of death are declining while Alzheimer's deaths are on the rise. Thus, knowing the causes of the disease helps find the best way to cure it. The goal of this present study is to develop a real data driven statistical predictive model using multiple logistic regression to predict Alzheimer's disease patients by selecting the relevant risk factors and their possible interactions using backward elimination then rank them based on their relative importance. By defining and ranking the statistically significant risk factors, they may be useful as a screening tool to discriminate Alzheimer's disease patients from cognitively normal individuals.

Multiple logistic regression is a type of probabilistic statistical classification model that is used to predict a binary dependent variable based on more than one predictor variables and their interactions. In

this model, there is a logistic transformation of the odds (logit) that will serve as the dependent variable. The odds are denoted as

$$\text{odds} = \frac{p}{1-p} \in (0, \infty),$$

and the multiple logistic regression model is defined as

$$\text{logit}[p] = \log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j$$

where p is the probability of selecting Alzheimer's patient, $\beta_j's$ indicate the coefficients (weights), and $X's$ are the risk factors.

1.2.2 Alzheimer's: Probabilistic Approach to the Behavior of Beta-Amyloid and Phosphorylated Tau Proteins Levels

Parametric Analysis is a statistical methodology that involves identification of a probability distribution function that characterizes the behavior of a given phenomenon of interest. Based on the identified probability distribution function, the maximum likelihood estimates of the parameters are obtained along with an appropriate degree of confidence. First, in this study, we have identified the probability distribution of phosphorylated tau and amyloid beta proteins that are collected and measured in pictogram/milliliter from the cerebrospinal fluid (CSF) of Alzheimer's patients. The cerebrospinal fluid is a clear fluid that surrounds the brain and spinal canal, which physicians can sample through a procedure called a lumbar puncture. Since it is in direct contact with the brain, any changes in cerebrospinal fluid (CSF) biomarkers are representative of changes in the brain.

When fitted the statistical distribution that best characterizes the behavior of the two proteins, we used commonly used goodness of fit tests namely: Kolmogorov-Smirnov, Anderson Darling, and Chi-squared. Kolmogorov Smirnov is based on minimum difference estimation. Anderson-Darling measures whether the data can transform into the uniform probability distribution. The Chi-square test for goodness of fit is a measure of relative error squared.

The null hypothesis for all goodness of fit test is that the data follow the desired probability distribution, and the alternative hypothesis is that the data does not follow the assumed probability distribution. For the Anderson Darling goodness of fit test, the test statistic is:

$$A^2 = -N - S,$$

where

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\ln F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right]$$

with F being the cumulative distribution function and Y_i the actual ordered data.

For the Kolmogorov-Smirnov goodness of fit test, the test statistics is defined as:

$$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right),$$

with F as the cumulative distribution function and Y_i the actual ordered data.

For the Chi-squared goodness of fit test, the test statistics is defined by

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i,$$

where $O_i$is the observed value and E_i is the expected value of the data based on the assumed probability distribution. As in the hypothesis testing, we fail to reject the null hypothesis and conclude that the assumed distribution ft our data well if the test statistics of the above tests is less than the critical value.

Second, we try to understand the bivariate probabilistic behavior of phosphorylated tau and beta-amyloid proteins in the spinal fluid that as their pattern indicates the likelihood of Alzheimer's disease. Thus, we model the joint behavior of phosphorylated tau and beta-amyloid levels by developing their bivariate probability distribution with their identified marginals and certain degrees of correlation. We used the copula method that links marginal probability distributions together to form a joint probability distribution. We found that 90-degree rotated Joe-Frank (BB8) copula function is the best copula function that fits our data. Having such a bivariate probability distribution, we can calculate different characterization of their bivariate behavior and finding a drug that can control their levels. Controlling their

level may help discover an effective treatment for Alzheimer's disease as scientists believe that they are an essential marker of this severe disease.

Chapter Two

Statistical Analysis and Modelling of the Atmospheric Carbon Dioxide in the Middle East and Comparisons with USA, EU, and South Korea

Note to Reader

This Chapter has been previously published in the Journal of Environment Vol. 1, No. 2, and have been reproduced with permission from SCIREA publishing [2].

2.1 Introduction

Global warming is a critical issue that our planet is facing and touching every part of the world. It is the increase in average global temperatures that is caused mostly by increased carbon dioxide (CO_2). The question is, how is carbon dioxide connected to global warming? Carbon Dioxide is present in the earth's atmosphere in a minimal amount, but it has a significant impact on life sustainability on the planet. It plays a crucial role in trapping heat in the atmosphere and keeping our world from freezing. However, the way that humans live, burning fossil fuels and other practices that release (CO_2) into the air, contributes to the amount of atmospheric carbon dioxide as CO_2 concentrations in Earth's atmosphere continue to increase, which raises the temperature of the planet. Thus, A warmer atmosphere means changes in normal climate patterns.

During the last two decades, CO_2 emissions in the Middle East countries have increased by over 200% based on The Energy Information Administration [3]. For this reason, in this chapter, we first performed a parametric statistical analysis to understand the probabilistic behavior of the atmospheric carbon dioxide in the Middle East. Then, we developed a data-driven nonlinear statistical model to identify

the significant attributable variables and all possible interactions and high order terms if applicable. Individual variables with significant interactions are ranked based on their percentage of contribution to CO_2 in the atmosphere and compared with those of the United States, European Union, and South Korea. Our statistical model has been evaluated by R squared (R^2), adjusted R squared (R^2_{adj}), and residual analysis.

Finally, the proposed statistical model will examine the major determinants that affect CO_2 in the atmosphere and illustrate different combinations of various attributable variables. Besides, this model will predict the atmospheric CO_2 given the information of the explanatory variables to suggest recommendations for these countries to reduce their CO_2 emissions level.

2.2 The Data

We used monthly data spanning from 1980-2008 of 15 countries in the Middle East, namely: Bahrain, Cyprus, Israel, Iran, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen, and the Occupied Palestinian Territory. The atmospheric CO_2 (parts per million) and CO_2 Emission (thousand metric tons of carbon) were obtained from the Carbon Dioxide Information Center (CDIC) [4], [5]. CO_2 emissions are byproducts of burning fossil fuels (gas fuels, liquid fuels, and solid fuels) manufacturing cement, and gas flares.

We used the average of two sampling sites to gather the data of total CO_2 in the atmosphere: Negev Desert, Israel, and Seychelles on Mahe Island. Israel's place is the only measurement site in the Middle East. Due to this data limitation, we included Seychelles' data as well. We used this site because of its location in the Indian Ocean and the effect of the ocean current making the data of Mahe Island partly representative. A map of the Middle East countries (shaded green) with measurement sites (red pins) is shown in Figure 2.1.

Moreover, from an economic point of view, many studies have indicated that there is a connection between CO_2 emissions and economic growth. For example, Farhani and Ben Rejeb [6] conducted a study to examine the relationship between energy consumption, economic growth (GDP), and CO_2 emissions. They found that an increase in energy consumption might lead to a rise in income and CO_2 emissions. Also, Al-Mulali [3] examined the relationship between CO_2 emission with energy consumption, economic growth, total exports and imports of goods and services, and foreign direct investment net inflows, which revealed that The total primary energy consumption, foreign direct investment net inflows, GDP, and total trade were essential factors in increasing CO_2 emissions. In this study, however, we focus only on the significance of the different types of fossil fuels, cement, and gas flares contributing to atmospheric CO_2 in the Middle East.

Figure 2. 1 Map of the Middle East countries and the two measurement sites

2.3 Methodology

2.3.1 Parametric Analysis

Parametric Analysis is a statistical methodology to identify the probability distribution function that characterizes the behavior of the variables of interest and approximate parameters estimate. Having defined such a function, we can obtain useful information such as the expected value and confidence interval [7]. First, we show that the natural phenomena such as atmospheric CO_2 do not follow Normal distribution which is clearly expressed in the non-symmetry histogram in Figure 2.2. A p-value = 6.69e-09 of the Anderson-Darling normality test is compatible with the plot that the subject data's distribution is not Gaussian.

Second, to identify the probability distribution function (pdf) that best fit the atmospheric carbon dioxide in the Middle East, we used three standard statistical tests: Kolmogorov-Smirnov [8], Anderson-Darling [9] and Chi-square goodness of fit test [10]. Thus, we found that the Johnson SB probability distribution best characterizes the probabilistic behavior of CO_2 in the atmosphere in the Middle East.

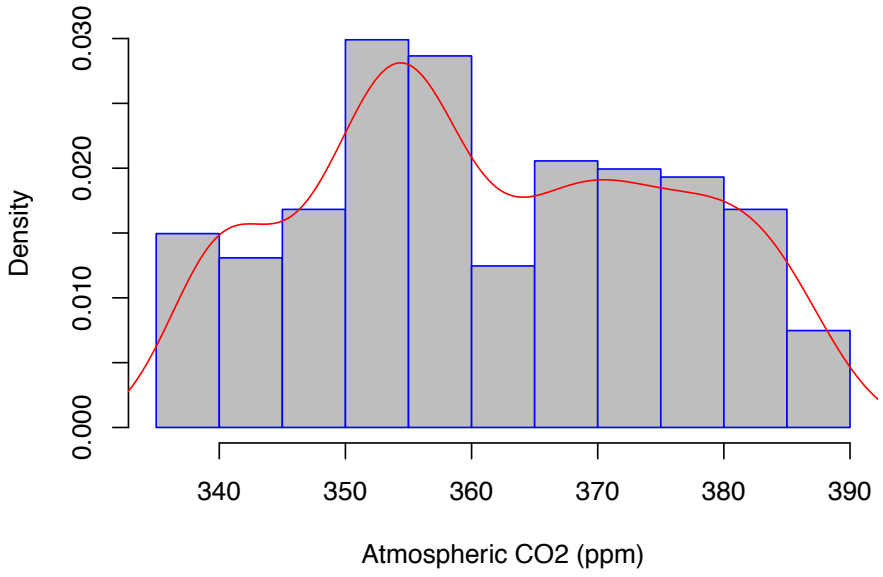


Figure 2. 2 Histogram of the atmospheric CO_2 in the Middle East

The probability density function of the Johnson SB distribution is given by:

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}\, z(1-z)}\, exp\left(-\frac{1}{2}\left(\gamma + \delta\, ln\left(\frac{z}{1-z}\right)\right)^2\right), \qquad \xi \le x \le \xi + \lambda \tag{2. 1}$$

Where $\gamma > 0$, $\delta > 0$ are the shape parameters, ξ is the location parameter, $\lambda > 0$ is the scale parameter and $z = \frac{x-\xi}{\lambda}$. The corresponding cumulative distribution function is given by

$$F(x) = \Phi\left(\gamma + \delta\, ln\left(\frac{z}{1-z}\right)\right) \tag{2. 2}$$

The approximate estimates of the four parameters γ, δ, λ and ξ of the Johnson SB probability distribution using the method of moments are given in Table 2.1.

Table 2. 1 Approximate parameters estimate of the Johnson SB distribution

Parameters	Approximate estimate
$\widehat{\boldsymbol{\gamma}}$	0.1189
$\widehat{\boldsymbol{\delta}}$	0.7372
$\widehat{\boldsymbol{\lambda}}$	55.479
$\widehat{\boldsymbol{\xi}}$	335.44

Thus, the pdf of Johnson SB probability distribution for the atmospheric carbon dioxide in the Middle East with $55.479 \le x \le 390.919$ is given by

$$f(x) = \frac{(0.29)exp\left(-\frac{1}{2}\left(0.1189 + 0.7372\, ln\left(\frac{0.003\,(-55.479+x)}{1-0.003(-55.479+x)}\right)\right)^2\right)}{\big(1-0.003(-55.479+x)\big)(-55.479+x)} \tag{2. 3}$$

We can use the cumulative distribution function to calculate the probability that a randomly chosen month has an atmospheric carbon dioxide less than or equal certain value. That is, $P(X < 355) \approx 0.3709$ is the probability that a randomly chosen month has an atmospheric carbon dioxide less than or equal to 355 ppm is approximately 0.3709. The graph of the pdf and CDF of the atmospheric CO_2 in the Middle East are given in Figure 2.3 and Figure 2.4, respectively.

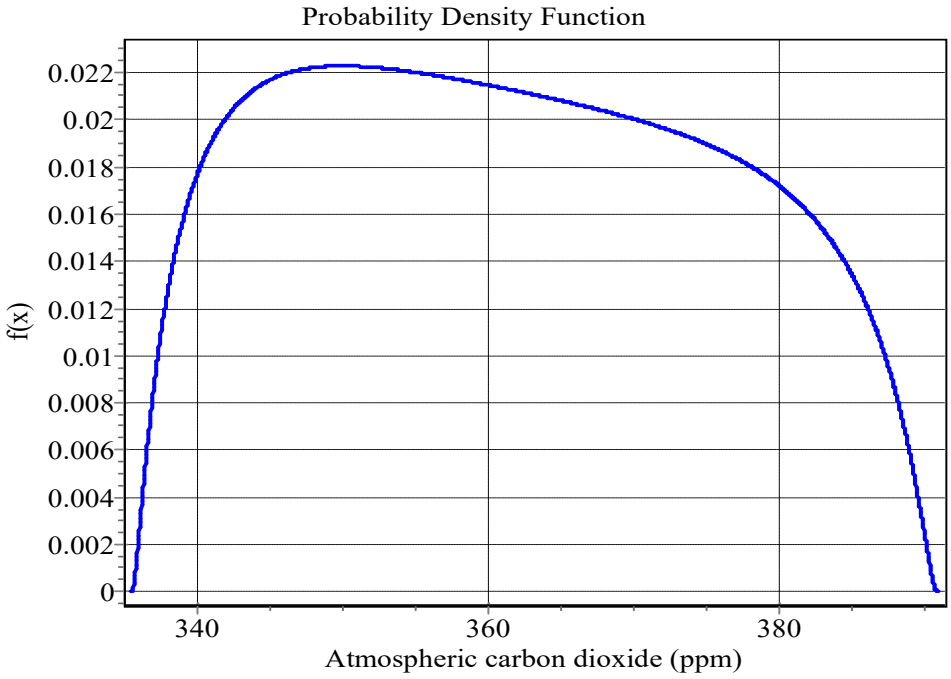


Figure 2. 3 Probability distribution function of the atmospheric CO_2

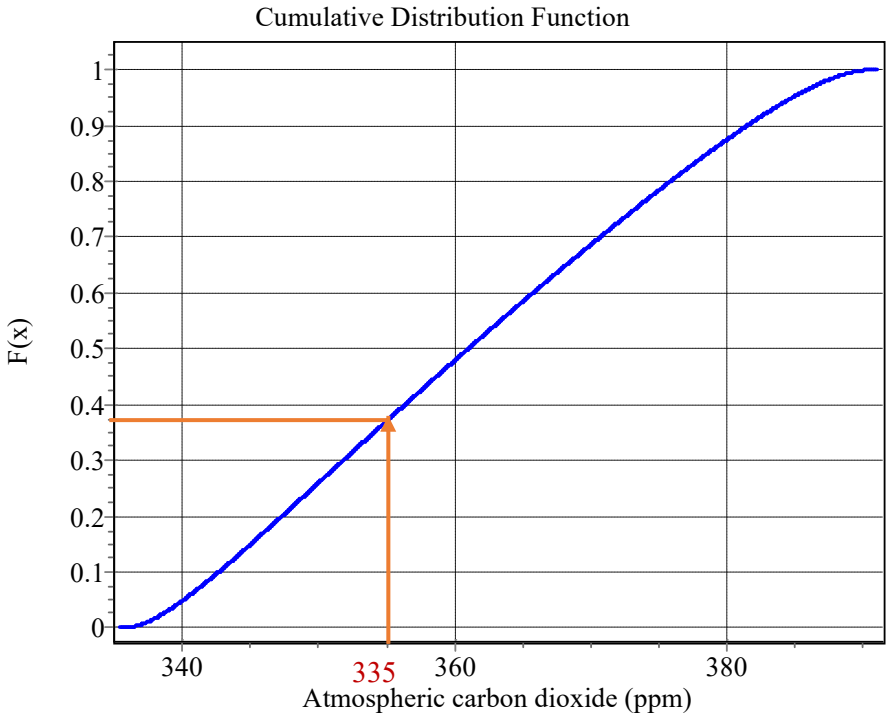


Figure 2. 4 Cumulative distribution function of the atmospheric CO_2

2.3.2 Non-linear Statistical Model

To achieve the goal of this study, a data-driven non-linear statistical model of CO_2 in the atmosphere was designed considering the atmospheric CO_2 as the response variable and gas fuels (Ga), solid fuels (So), liquid fuels (Li), gas Flares (Fl) and cement (Ce) as explanatory variables. Therefore, the statistical form of the model with all possible interactions is:

$$CO_2 = \beta_O + \beta_1 Ga + \beta_2 So + \beta_3 Li + \beta_4 Fl + \beta_5 Ce + \beta_6 A_1 + \cdots + \beta_j A_j + \varepsilon\,, \qquad (2.4)$$

where CO_2 Indicates the atmospheric CO_2, $\beta' s$ are the coefficients, $A' s$ are all possible interactions and high order terms and ε is a random error. The assumption to build the above model (2.4) is that the response (atmospheric CO_2) should follow Gaussian distribution and we have proven that it does not follow Normal. Thus, we apply Johnson transformation to the atmospheric CO_2, which results in the following equation,

$$Y_t = 0.0851 + 0.7302 \ln\left[\frac{CO_2 - 335.0911}{390.7866 - CO_2}\right], \qquad (2.5)$$

where Y_t is the transformed data as close to Gaussian distribution as possible. After satisfying the normality distribution, we develop our statistical model by starting with the full model that included all five attributable variables and all possible interaction terms. Using backward elimination, we found that only three out of five explanatory variables are significantly contributing to CO_2 in the atmosphere with only five interaction terms. Thus, the best statistical model with all the significant factors and interactions that influence CO_2 in the atmosphere in the Middle East is given by:

$$\begin{aligned}\widehat{Y_t} = {} & -2.11 + 2.121 * 10^{-5}\, Ga - 1.041 * 10^{-5} Li + 1.323 * 10^{-4} Ce \\ & -1.34 * 10^{-9}\, Ga\, So + 1.082 * 10^{-9}\, Ga\, Ce + 4.07 * 10^{-10}\, Li\, So \\ & +4.668 * 10^{-9}\, So\, Fl - 6.868 * 10^{-9} Fl\, Ce.\end{aligned} \qquad (2.6)$$

In accordance with the proposed statistical model, gas fuels, liquid fuels, and cement are identified as key factors contributing to CO_2 in the atmosphere in the Middle East. Furthermore, the statistical model

identified the following interactions that are statistically significant to the atmospheric CO_2 namely (Gas Fuels*Solid Fuels), (Gas Fuels*Cement), (Liquid Fuels*Solid Fuels), (Solid Fuels*Gas Flares) and (Gas Flares*Cement).

The estimated $\widehat{Y_t}$ from equation (2.6) is based on the transformed data. Thus, we can get the estimated value of the atmospheric carbon dioxide CO_2 from

$$\widehat{CO_2} = \frac{335.099 + 347.706\, e^{0.7302\, Y_t}}{1 + 0.8897\, e^{0.7302\, Y_t}}. \qquad (2.\ 7)$$

The recommended statistical model has been assessed using R squared (R^2) and adjusted R squared (R^2_{adj}) which are the key criteria to evaluate the model fitting. They provide an overall measurement of how well the model fits. The regression sum of squares (SSR), is the variation that is explained by the proposed model. The sum of squared errors (SSE), known as the residual sum of squares, is the variation that is left unexplained. The total sum of squares (SST) is proportional to the sample variance and equals the sum of SSR and SSE[11]. The coefficient of determination (R^2) represents the proportion of total variation in the response that is explained by the proposed statistical model and is given by

$$R^2 = 1 - \frac{SEE}{SST} \qquad (2.\ 8)$$

As (R^2) always increases with every explanatory variable added to a statistical model, adjusted R squared (R^2_{adj}) has been adjusted for the number of predictors in the model as follows: it increases only if more variables are added and improve the model. On the other hand, it decreases when we add more useless predictors to the model. It is preferred when we work with several parameters and is given by

$$R^2_{adj} = 1 - \frac{SEE/df_{error}}{SST/df_{total}} \qquad (2.\ 9)$$

For our proposed statistical model, R^2 is 0.9784 and R^2_{adj} is 0.9778. That is, our statistical model explains 97.87% of the variation in the response variable; equivalently, the significant attributable variables and the interactions estimate about 97% of the total atmospheric CO_2 in the Middle East. Both $R^2$and $R^2_{adj}$are very high (more than 90%) and very close to each other. These results illustrate that the increase

of the value of R^2 is not due to the increase in the number of the predictors but to the good quality of the proposed statistical model.

Additionally, we performed a residual analysis that calculated the actual value of CO_2 in the atmosphere (response) minus the estimated value of CO_2 in the atmosphere using the proposed statistical model, and it attests the quality of the developed statistical model. The residual analysis also justified model assumptions of normality, linearity, and constant error variance. For the developed statistical model, the mean residual was very small ($\bar{r} = \frac{1}{n}\sum r_i = 5.045 \times 10^{-18}$), and it indicates that the predictions from our statistical model are good. Moreover, the residual plots are used to assess the model assumptions such as Q-Q plots in Figure 2.5 and the scatter plot in Figure 2.6. In the Q-Q plot, we see approximate normality distributed residual, and the scatter plot illustrates an approximate zero mean and no clear pattern or trend in the residuals.

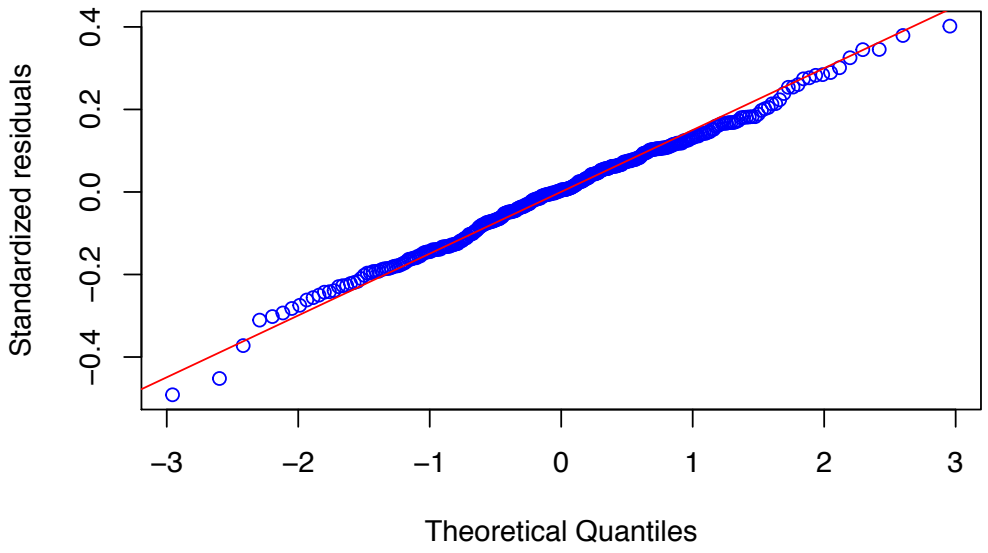


Figure 2. 5 Residual's Q-Q plot

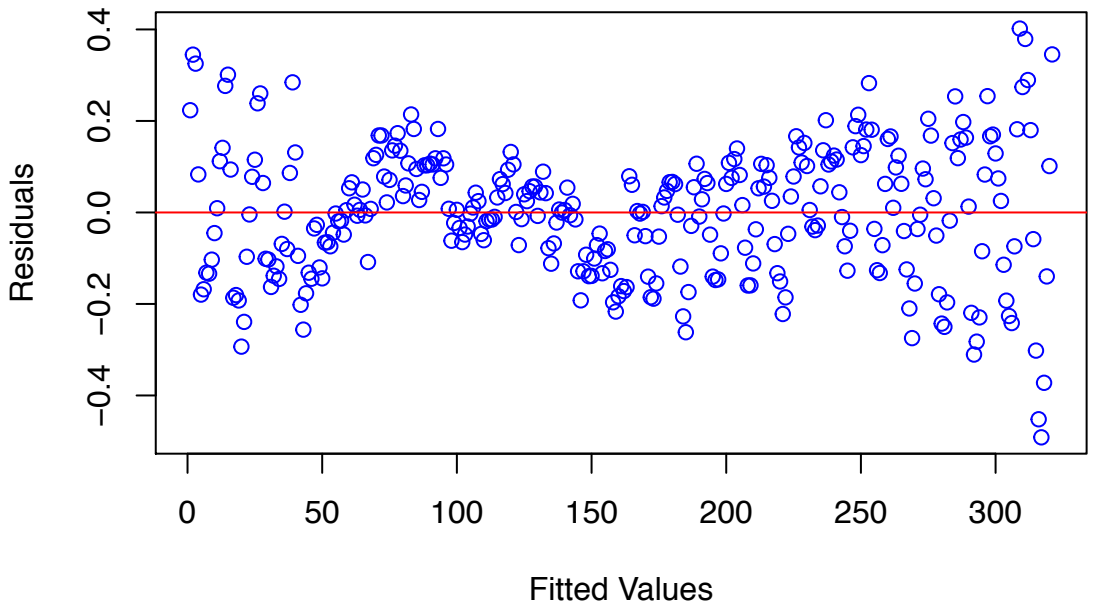


Figure 2. 6 Residual's scatter plot

2.4 Results and Discussion

After obtaining the proper statistical model and evaluating the quality of the model using different criteria, which are stated above, we can infer useful pieces of information from the subject model. First, we will identify the significant attributable variables and their interaction terms. That is, we can identify gas fuels, liquid fuels, cement and the interactions of (gas fuels* solid fuels), (gas fuels*cement), (liquid fuels*solid fuels), (solid Fuels*gas flare), and (Gas flare * Cement) as the key factors affecting CO_2 in the atmosphere. Second, we can use the model to predict the atmospheric CO_2 given the information of the attributable variables and pose recommendations for these countries to reduce their CO_2 emissions level.

Third, one of the advantages of the proposed statistical model is to rank the variables and their significant interactions based on their percentage of contribution to CO_2 in the atmosphere [12]. As seen in Table 2.2, cement manufacturing is ranked as the 1st contributing predictor to the atmospheric CO_2 in the Middle East which contribute to about 15.28% to the CO_2. The next most significant contribution is gas fuels with about 14.70%. Also, the interaction between the gas flare and cement production has the lowest

rank with about 7.90% of contribution to the atmospheric CO_2 in the Middle East. In Figure 2.7, we ranked the significant risk factors and their interactions by their percentage of contribution to the CO_2 in the atmosphere in the Middle East.

Fourth, we can perform a surface response analysis to identify the value of each contributable variable to minimize the CO_2 emissions in the atmosphere. Finally, we can calculate the confidence limit, which will be useful in controlling CO_2 emission.

Table 2. 2 Ranking the Variables Based on their Contribution

Rank	Variables	Contribution (%)
1	Cement Production	15.28
2	Gas Fuels	14.70
3	Liquid Fuels* Solid Fuels	13.66
4	Gas Fuels* Solid Fuels	13.47
5	Gas Fuels* Cement	12.56
6	Liquid Fuels	10.63
7	Solid Fuels* Gas Flare	9.65
8	Gas Flare* Cement	7.90

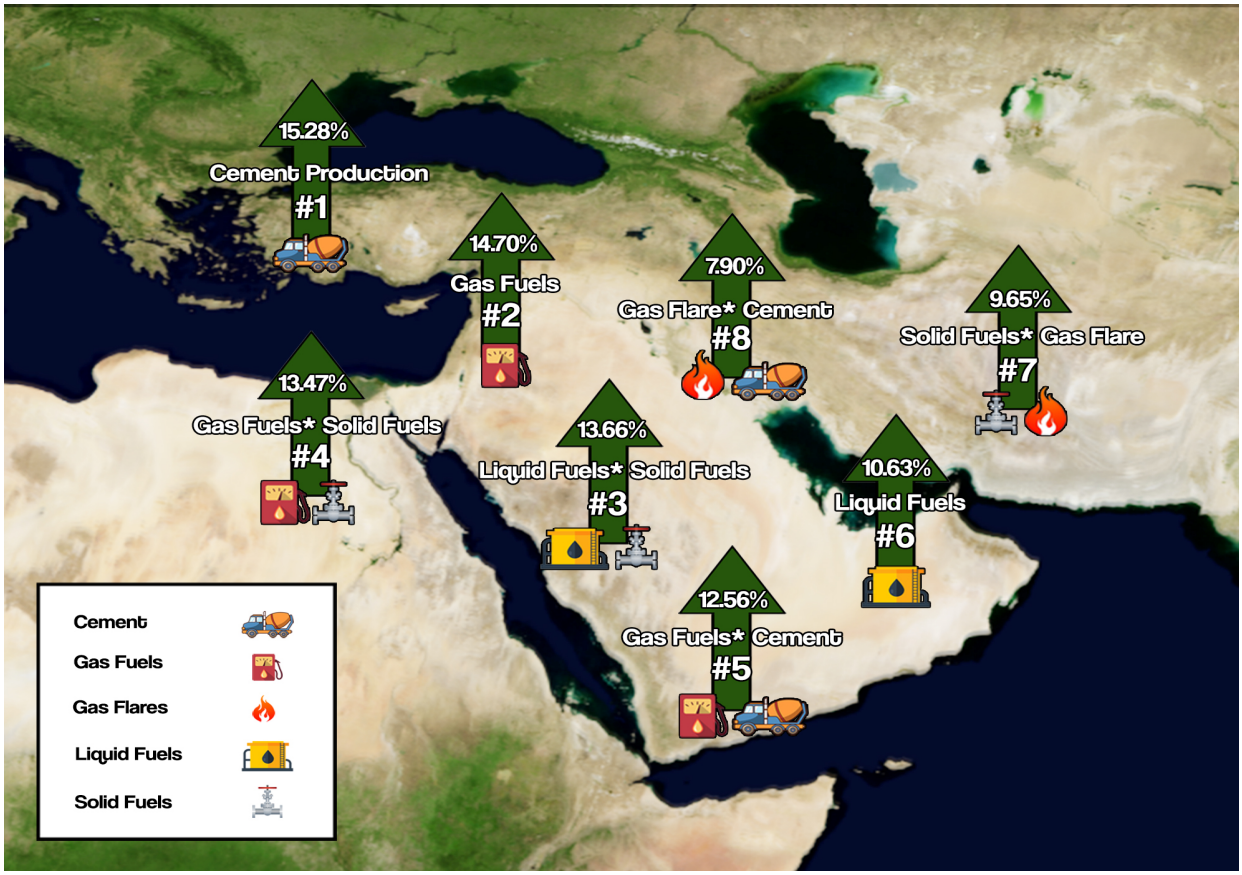


Figure 2. 7 Ranking of the attributable variables contributing to the atmospheric CO_2 in the Middle East

*Map source: https://svs.gsfc.nasa.gov/vis/a000000/a003400/a003487/earth4K.png
*Modified by Abdul-Aziz Habadi

2.5 Comparison between the USA, EU, South Korea, and the Middle East

Since world leaders agree that global warming is a serious problem, there is more international consensus to establish a global policy to control the factors of global warming. To support this idea, we will do a comparative analysis of the atmospheric carbon dioxide between the USA, EU, South Korea, and the Middle East.

Xu and Tsokos (2013) [13] built a statistical model that identified the significant risk factors and their interactions that contribute to the CO_2 in the atmosphere in the United States. These variables and

interactions contributed to about 98.98% of CO_2 emissions in the United States. The ranks of the contributing variables with the rate of contribution of CO_2 in the atmosphere are listed in Table 2.3

Similarly, Teodorescu and Tsokos (2013) [14] structured a statistical model using CO_2 emissions data for countries within the European Union (EU). They found that gas-fuels create about 48.72% of overall CO_2 emissions in the EU. The significant risk factors and their interactions along with their ranking are presented in Table 2.4 below.

.

Table 2. 3 Ranking the Attributing Variables of USA

Rank	Variables	Contribution (%)
1	Liquid-Fuels (Li)	17.59
2	Li & Ce	16.36
3	Ce & Bu	15.73
4	Bunker-Fuels (Bu)	15.06
5	Cement (Ce)	10.77
6	Gas-Flares (Fl)	8.95
7	Gas-Fuels (Ga)	6.82
8	Ga & Fl	5.43
9	Li & Ga	2.25
10	Li & Bu	0.02

Table 2. 4 Ranking the attributable variables of EU

Rank	Variables	Contribution (%)
1	Gas-Fuels (Ga)	48.72
2	Li & Bu	12.41
3	Li2	11.79
4	Bu2	7.78
5	Gas-Flares (Fl)	6.66
6	Li & Fl	5.06
7	Li & Bu	4.71
8	Liquid-Fuels (Li)	2.86

Recently, Kim and Tsokos (2015) [15] have identified the individual attributable variables along with significant interactions terms that contribute to atmospheric CO_2 in South Korea. Their proposed statistical model explained 99.41% of the CO_2 in the atmosphere. The ranking of the explanatory variables and significant interactions with their percentages of overall contribution are presented in Table 2.5.

Also, Table 2.6 gives an interesting comparison of what contributes to the CO_2 in the atmosphere in the United States, European Union, South Korea, and the Middle East. A significant fact we get from this comparison is that the most massive CO_2 emission in the Middle East is caused by cement productions, whereas in the US is the 5th contributing variable to the atmospheric CO_2. Moreover, liquid fuels are ranked as the number one attributable variable in the US and South Korea; however, it is the 6th in the Middle East with 10.63% contribution to the atmospheric carbon dioxide.

Also, the Middle East and the US have five significant interactions of the risk factors while South Korea has six, and the EU has only three contributing interactions to CO_2 emissions. These comparisons support the idea that each country should form its policy to regulate this issue individually.

Table 2. 5 Ranking the attributable variables of South Korea

Rank	Variables	Contribution (%)
1	Liquid-Fuels (Li)	75.37
2	Solid-Fuels (So)	18.61
3	So & Bu	2.008
4	Ga & Bu	1.534
5	Li & Bu	0.912
6	Bunker-Fuels (Bu)	0.47
7	Gas-Fuels (Ga)	0.224
8	Li & So	0.207
9	Li & Ga	0.062
10	Li & So & Bu	0.004

Table 2. 6 Comparison between the USA, the EU, South Korea, and ME

Rank	USA	South Korea	EU	Middle East
1	Li	Li	Ga	Ce
2	Li & Ce	So	Li & Bu	Ga
3	Ce & Bu	So & Bu	Li^2	Li* So
4	Bu	Ga & Bu	Bu^2	Ga* So
5	Ce	Li & Bu	Fl	Ga* Ce
6	Fl	Bu	Li & Fl	Li
7	Ga	Ga	Li & Bu	So* Fl
8	Ga & Fl	Li & So	Li	Fl* Ce
9	Li & Ga	Li & Ga	-	
10	Li & Bu	Li & So & Bu	-	

2.6 Conclusion and Contributions

First**,** we performed a parametric analysis of the atmospheric CO_2 in the Middle East and found that Johnson SB probability distribution best characterizes the probabilistic behavior of this natural phenomenon. Second, we developed a data-driven non-liner statistical model that identifies the risk factors and their interaction terms that affect the atmospheric CO_2 in the Middle East. We have found that gas-fuels, liquid fuels, cement, and only five interaction terms namely (gas fuels* solid fuels), (gas fuels*cement), (liquid fuels*solid fuels), (solid Fuels*gas flare), and (Gas flare * Cement) are significantly contributing to atmospheric CO_2. The proposed statistical model was evaluated using R squared (R^2), adjusted R squared (R^2_{adj}) and residual analysis. All the results supported the high quality of our proposed statistical model.

Several significant points can be obtained from our proposed statistical model. First, this model can be used to get an accurate estimate of CO_2 in the atmosphere. Second, it can be used to identify the significant attributable variables and their interaction terms and rank them based on their percentage of contribution to CO_2 in the atmosphere. Finally, we can utilize surface response analysis to identify the value of the contributable variables and interaction that will help to develop a strategic policy to control or minimize CO_2 emissions in the Middle East.

Moreover, we have compared the predictors of the atmospheric CO_2 in the Middle East with those of the United States, European Union countries, and South Korea. Some of the interesting comparisons are: cement productions are the number one factor of the CO_2 emissions in the Middle East and contribute about 15.28%, wherein the US is the number five and contributes about 10.77%. Also, liquid fuels ranked as the number one attributable variable in the US and South Korea; however, it is the 6th in the Middle East with 10.63% contribution.

The results of this study lead us to conclude that there is no need for a global policy to control global warming, but each country should establish its policy to regulate this issue individually.

Our contributions to this chapter can be summarized as follows:

1. We identified that the Johnson SB probability distribution best characterizes the probabilistic behavior of the atmospheric CO_2 in the Middle East.
2. We developed a data-driven statistical non-linear model that identifies the risk factors and their interaction terms that affect the atmospheric CO_2 in the Middle East.
3. Evaluate the quality of our proposed model using R^2, R^2_{adj}, and residual analysis
4. We ranked the significant risk factors based on their percentage of contribution to CO_2 in the atmosphere.
5. We compared the results of our model with the finding of the United States, the European Union, and South Korea.
6. The developing model and the comparison are useful in structuring regional strategic policies and plans, but not global, to maintain an optimal level of CO_2 in the atmosphere.

Chapter Three

Statistical Forecasting Models of Atmospheric Carbon Dioxide and Temperature in the Middle East

Note to Reader

This Chapter has been previously published in journal of Geoscience and Environment Protection Vol. 5 No.10, and have been reproduced with permission from Scientific Research Publishing [16].

3.1 Introduction

Time series analysis is an interesting and important statistical procedure that can be used for forecasting the phenomenon of interest. This statistical method depends on tracking the phenomena (or variable) over a given time period and then predict the future based on the different values in the time series and on the pattern of growth in values. The aim of the present study is to develop statistical time series forecasting models to predict carbon dioxide (CO_2) in the atmosphere in the Middle East and atmospheric temperature in Saudi Arabia.

Since it is well known that the most fundamental cause of global warming is the excessive rise of greenhouse gasses, probably the product of the industrial revolution, that accumulate in the atmosphere, blocking heat and leading to increased temperatures within the Earth's atmosphere. Especially, the raise proportion of the carbon dioxide from their very normal level has the most significant effect on substantial changes in the Earth's climate. The Middle East is emitting approximately 1,714.09 million metric tons of carbon dioxide into the atmosphere, and based on U.S department of energy, three Middle Eastern countries are among the five highest national per capita CO_2 emissions rates in the world for 2008: Qatar (14.58 metric tons of carbon per person), United Arab Emirates (9.43), and Bahrain (7.90) [17].

In chapter two, we have developed a statistical model that identifies the risk factors of the atmospheric CO_2 in the Middle East affected by carbon dioxide emission that is related to fossil fuels, gas flares, cement production, and their interaction terms. We have found that gas-fuels, liquid fuels, cement, and only 4 interaction terms namely (Liquid Fuels*Solid Fuels), (Liquid Fuels*Gas Flares), (Solid Fuels* Cement) and (Gas Flares * Cement) are significantly contributing to atmospheric CO_2 in the Middle East and we compare our statistical finding with the statistical models of the atmospheric carbon dioxide in the United states, Europe and South Korea [2], [15].

Thus, the objective of the present study is to develop two different statistical time series forecasting models for the atmospheric carbon dioxide concentration in the Middle East, in addition to the atmospheric temperature in Saudi Arabia. These two forecasting models are useful in monitoring the future level of carbon dioxide emission in the Middle East.

3.2 Atmospheric CO_2 Statistical Forecasting Model

To develop our statistical forecasting model, we used monthly data of atmospheric carbon dioxide concentrations measured in part per million from 1996 to 2015. The data was collected in Weizmann Institute of science at the Arava Institute and provided by National Oceanic and Atmospheric Administration, Earth system research laboratory, Global Monitoring Division, Boulder, Colorado, USA (http://esrl.noaa.gov/gmd/). Figure 3.1 below gives a visual presentation of the time series plot of atmospheric CO_2 in the Middle East.

The data is clearly non-stationary with seasonality and increasing trend. Most of the time series we encounter in real world problems are non-stationary, and we must remove non-stationary component to utilize methodology for stationary time series data. Thus, in order for us to do the analysis, we must first reduce a non-stationary time series into a stationary time series after applying a proper degree of difference filter of the given series. Since we have a seasonal data, the multiplicative seasonal autoregressive integrated

moving average (seasonal ARIMA) model will be used to develop the statistical predictive model of the atmospheric carbon dioxide in the Middle East[18]–[20].

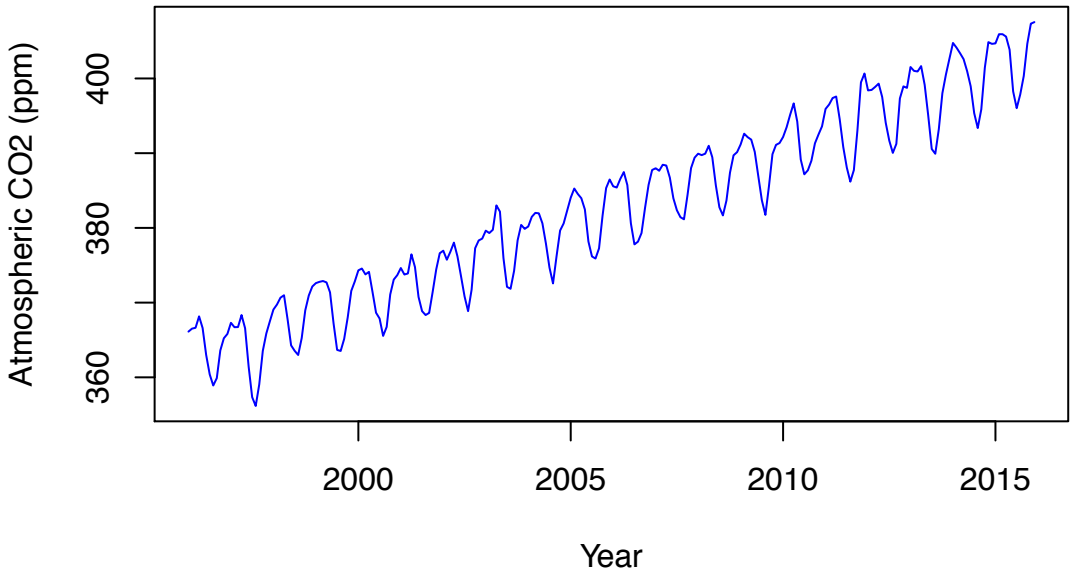


Figure 3. 1 Time Series plot of the atmospheric CO_2 data of the Middle East from 1996-2015

A seasonal ARIMA model is formed by including seasonal terms in the autoregressive integrated moving average model $ARIMA\ (p, d, q)$ as is defined as follows

$$\phi_{\mathrm{p}}(\mathrm{B})(1-\mathrm{B})^{\mathrm{d}}\mathrm{x}_{\mathrm{t}} = \theta_{\mathrm{q}}(\mathrm{B})\varepsilon_{\mathrm{t}}\,, \tag{3.1}$$

where p is order of autoregressive process, d is degree of differencing (filter); q is order of moving average, and the analytical form of seasonal ARIMA $(p, d, q)(P, D, Q)_s$ is defined by

$$\Phi_{\mathrm{P}}(B^S)\ \phi_p(B)\ (1-B)^d(1-B^S)^D x_t = \theta_q(B)\ \Theta_Q(B^S)\varepsilon_t\,, \tag{3.2}$$

where p, d and q as defined above, also , P is the order of the seasonal autoregressive process, D is the order of the seasonal differencing, Q is the order of the seasonal moving average, and the subindex S refers to the seasonal period, with monthly data S=12; for quarterly data S=4, and $\Phi_{\mathrm{P}}(B^S), \phi_p(B), \theta_q(B), \Theta_Q(B^S)$ are defined as follows:

The non-seasonal components we have:

$$\text{AR: } \phi_p(B) = \left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)$$

and

$$\text{MA: } \theta_q(B) = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right)$$

The seasonal components are:

$$\text{Seasonal AR: } \Phi_{\mathrm{P}}(B^S) = (1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS})$$

and

$$\text{Seasonal MA: } \Theta_Q(B^S) = \left(1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{QS}\right)$$

In the present study, since we have a monthly data, we let the seasonal subindex S=12. Once we transform our data into stationary time series, we found that the best statistical forecasting model that characterizes the monthly atmospheric carbon dioxide concentration in the Middle East with minimum AIC [21] is ARIMA $(2,1,3)(0,1,1)_{12}$; analytically is given by

$$(1 - \phi_1 \mathrm{B} - \phi_2 \mathrm{B}^2)(1 - \mathrm{B})(1 - \mathrm{B}^{12})\mathrm{x_t} = (1 + \theta_1 \mathrm{B} + \theta_2 \mathrm{B}^2 + \theta_3 \mathrm{B}^3)(1 + \Theta_1 \mathrm{B}^{12})\varepsilon_\mathrm{t} \qquad (3.3)$$

with first non-seasonal difference filter and first seasonal difference filter, second order of non-seasonal autoregressive process AR (2), third order of non-seasonal moving average process MA (3), and first order of seasonal moving average process SMA (1). Expanding both sides of the above ARIMA model, we have

$$\begin{gathered}[1 - (1 + \phi_1)B + (\phi_1 - \phi_2)B^2 + \phi_2 B^3 - B^{12} + (1 + \phi_1)B^{13} \\ +(\phi_2 - \phi_1)B^{14} - \phi_2 B^{15}]\, x_t = [1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 \\ +\Theta_1 B^{12} + \theta_1 \Theta_1 B^{13} + \theta_2 \Theta_1 B^{14} + \theta_3 \Theta_1 B^{15}]\varepsilon_t\end{gathered} \qquad (3.4)$$

Simplify it and using backshift operation $B^j x_t = x_{t-j}$, we obtain

$$\begin{gathered}x_t = (1 + \phi_1)x_{t-1} - (\phi_1 - \phi_2)x_{t-2} - \phi_2 x_{t-3} + x_{t-12} - (1 + \phi_1)x_{t-13} \\ -(\phi_2 - \phi_1)x_{t-14} + \phi_2 x_{t-15} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} \\ +\Theta_1 \varepsilon_{t-12} + \theta_1 \Theta_1 \varepsilon_{t-13} + \theta_2 \Theta_1 \varepsilon_{t-14} + \theta_3 \Theta_1 \varepsilon_{t-15}\end{gathered} \qquad (3.5)$$

Thus, the approximate maximum likelihood estimates of the coefficients are

$$\phi_1 = -0.6791, \phi_2 = 0.1376, \theta_1 = 0.9140$$

$$\theta_2 = -0.8964\,, \theta_3 = -0.8803, \Theta_1 = -0.9996\,,$$

by letting $\varepsilon_t = 0$, the one-step ahead forecasting model for atmospheric CO_2 in the Middle East is given by

$$\begin{aligned}\hat{x}_t = {} & 0.3209x_{t-1} + 0.8167x_{t-2} - 0.1376x_{t-3} + x_{t-12} - \\ & 0.3209x_{t-13} - 0.8167x_{t-14} + 0.1376x_{t-15} + 0.9140\varepsilon_{t-1} - \\ & 0.8964\varepsilon_{t-2} - 0.8803\varepsilon_{t-3} - 0.9996\varepsilon_{t-12} - 0.9136\varepsilon_{t-13} \\ & +0.8960\varepsilon_{t-14} + 0.8799\varepsilon_{t-15}\end{aligned} \tag{3.6}$$

Once we identify the forecasting model of the atmospheric carbon dioxide, we need to evaluate or validate our proposed model and illustrate the quality of model. In Figure 3.2 below presents the actual data with the forecasting values of the atmospheric carbon dioxide in the Middle East that obtained by our proposed statistical forecasting model. In addition, we perform residual analysis and calculate the residuals estimates $r_t = x_t - \hat{x}_t$; Figure 3.3 below shows the graphical result of the residual estimates.

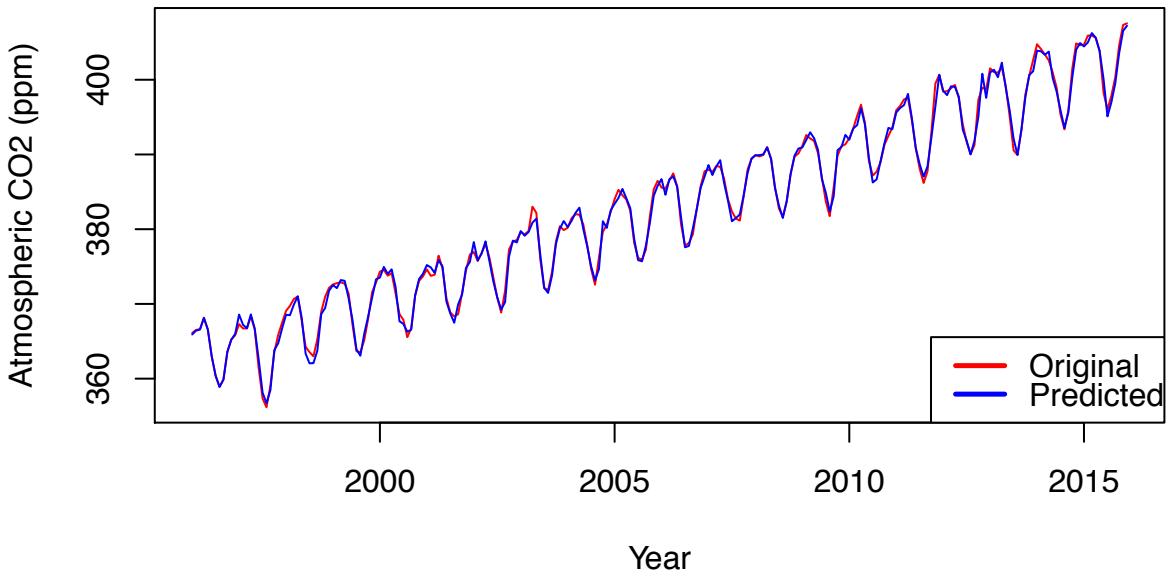


Figure 3. 2 Original vs. predicted values of the atmospheric CO_2

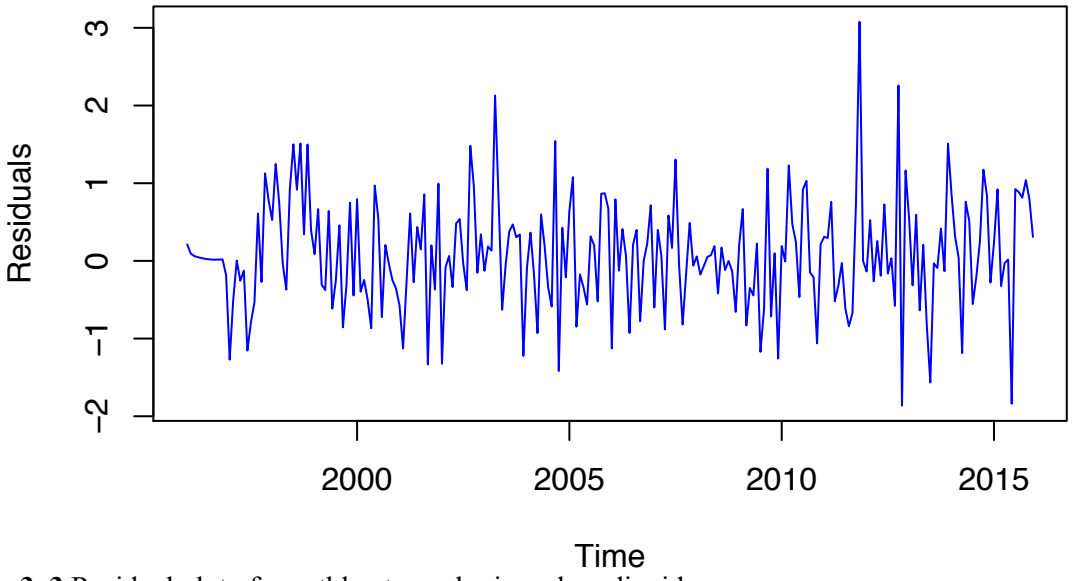


Figure 3. 3 Residual plot of monthly atmospheric carbon dioxide

As we can see in Figure 3.2 above, the predicted values follow the original data of the atmospheric CO_2. Furthermore, the residuals in Figure 3.3 are quite small and isolating around zero and that is an indication of the good quality of our proposed statistical time series-forecasting model of the atmospheric CO_2 in the Middle East.

Next, we evaluate the mean of the residuals, $\bar{r}$, the variance, S_r^2, and the mean square error, *MSE*, and the results are presented below in Table 3.1

Table 3. 1 Basic evaluation of the atmospheric carbon dioxide model

$\bar{r}$	S_r^2	*MSE*
0.0812	0.5062	0.5107

The results show the effectiveness of the proposed model for forecasting atmospheric carbon dioxide in the Middle East.

Furthermore, we restructure the model (3.6) with monthly data from 1996-2013 to forecast the last 24 hidden values of using the previous observations. The purpose is to test the accuracy of the forecasting values of the atmospheric CO_2 with respect to the observed 24 values that have not been used and how well the model performs on new data that were not used when fitting the model. Table 3.2 below gives the actual

and predicted values of carbon dioxide in the atmosphere.

Table 3. 2 Actual vs. forecasting values of atmospheric CO_2

Month	Original Values	Forecasting values	Residuals
Jan 2014	404.75	403.76	0.99
Feb 2014	404.12	402.60	1.52
Mar 2104	403.38	402.55	0.83
Apr 2104	402.58	403.45	-0.87
May 2104	400.97	401.43	-0.46
Jun 2104	398.95	397.63	1.32
Jul 2104	395.35	394.90	0.45
Aug 2104	393.36	393.97	-0.61
Sep 2104	395.86	395.76	0.10
Oct 2014	401.45	399.85	1.60
Nov 2014	404.86	402.25	2.61
Dec 2104	404.63	403.31	1.32
Jan 2015	404.69	404.16	0.53
Feb 2105	405.92	404.30	1.62
Mar 2015	405.92	404.60	1.32
Apr 2015	405.6	405.44	0.16
May 2015	403.84	403.51	0.33
Jun 2015	398.26	399.64	-1.38
Jul 2015	396.02	396.97	-0.95
Aug 2015	397.86	395.98	1.88
Sep 2105	400.33	397.83	2.50
Oct 2015	404.64	401.88	2.76
Nov 2015	407.33	404.31	3.02
Dec 2105	407.54	405.33	2.21

As we can see, the difference between the original and predicted values of the carbon dioxide in the Middle East is very small. Figure 3.4 gives a graphical presentation of the results in Table 3.2. Since the predicted values produced by our proposed statistical model are very close to the original values, and the forecast errors seem to be very small, the ARIMA $(2,1,3)(0,1,1)_{12}$does seem to provide an adequate predictive model for the atmospheric carbon dioxide in the Middle East.

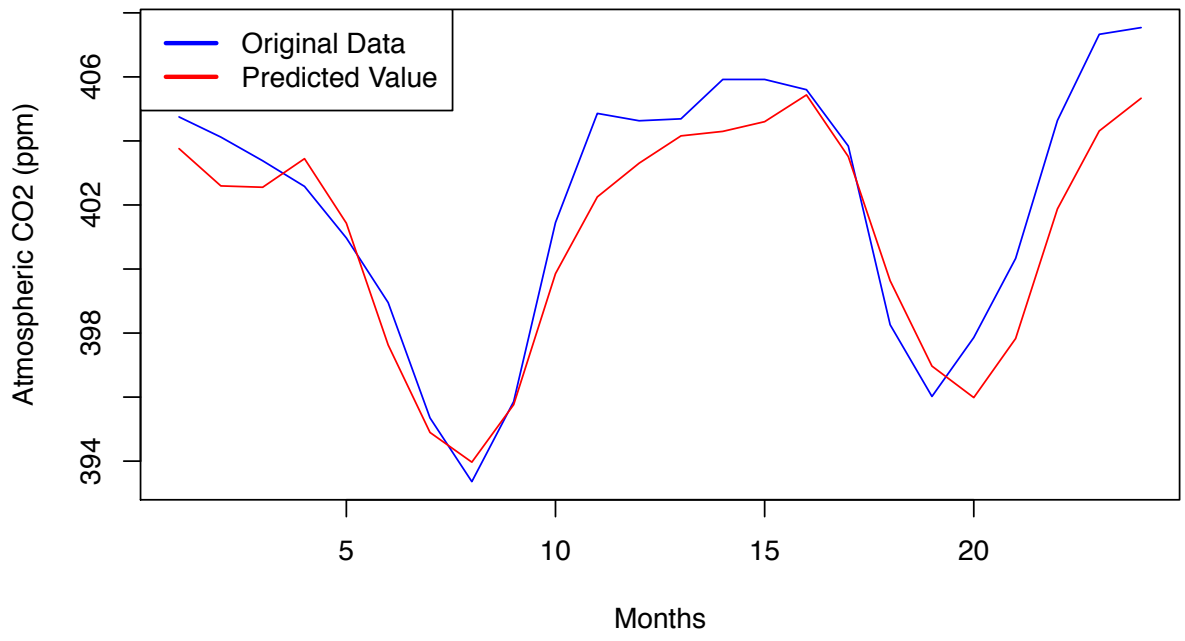


Figure 3. 4 Monthly atmospheric CO2 vs. predicted values of the last 24 months

3.3 Atmospheric Temperature Forecasting Model of Saudi Arabia

Saudi Arabia's prevailing climate is hot and dry, but according to weather expert, The Kingdom of Saudi Arabia has witnessed an unprecedented drop in temperature accompanied by uncommon natural phenomena. Frost and freezing temperatures and unusually heavy snowfall have been reported in several areas in Saudi Arabia in winter, as well as increasing the heat in summer. In general, the changes in the global climate due to the impact of global warming will lead to more extreme seasons. Thus, the aim of this part is to develop a statistical forecasting model for temperature in Saudi Arabia as temperature plays an important role in Global warming.

The dataset includes monthly average temperature measured in Celsius (°C) of Saudi Arabia as only available data from January 1970 to December 2015. The data was published by the Saudi's General Authority of Meteorology and Environmental protection. A presentation of the temperature data is given below in Figure 3.5.

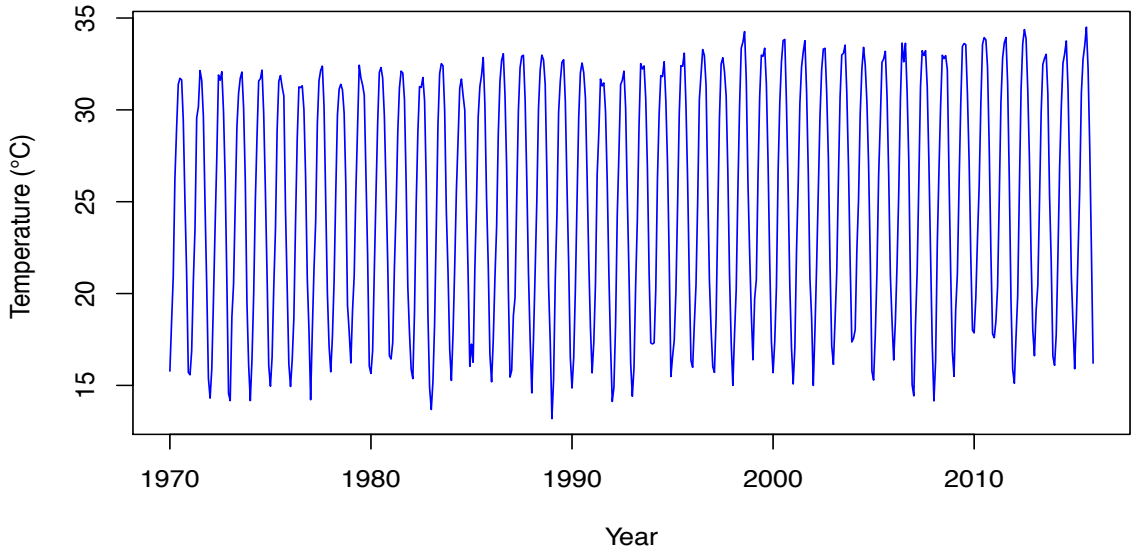


Figure 3. 5 Time series plot of monthly temperature of Saudi Arabia from 1970-2015

We will develop a forecasting model using the multiplicative seasonal autoregressive integrated moving average (seasonal ARIMA) model as described in section 3.2 [22], [23]. Thus, after confirming the stationary of our series and let the seasonal subindex S=12, we found the model that best described the monthly atmospheric temperature of the kingdom of Saudi Arabia is ARIMA $(1,1,2)(0,1,1)_{12}$, and analytically is given by

$$(1-\emptyset_1 \mathrm{B})(1-\mathrm{B})(1-\mathrm{B}^{12})\mathrm{x_t} = (1+\theta_1 \mathrm{B}+\theta_2 \mathrm{B}^2)(1+\Theta_1 \mathrm{B}^{12})\varepsilon_\mathrm{t} \tag{3.7}$$

with first non-seasonal difference filter and first seasonal difference filter, first order of non-seasonal autoregressive process AR (1), second order of non-seasonal moving average process MA (2), and first order of seasonal moving average process SMA (1). Expanding both sides, we have

$$\begin{aligned}&[1-(1+\emptyset_1)B+\emptyset_1 B^2 - B^{12}+(1+\emptyset_1)B^{13}-\emptyset_1 B^{14}]x_t \\ &= [1+\theta_1 B+\theta_2 B^2+\Theta_1 B^{12}+\theta_1\Theta_1 B^{13}+\theta_2\Theta_1 B^{14}]\varepsilon_t\end{aligned} \tag{3.8}$$

Simplify it, we get

$$\begin{aligned}&x_t-(1+\emptyset_1)x_{t-1}+\emptyset_1 x_{t-2}-x_{t-12}+(1+\emptyset_1)x_{t-13}-\emptyset_1 x_{t-14} \\ &= \varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\Theta_1\varepsilon_{t-12}+\theta_1\Theta_1\varepsilon_{t-13}+\theta_2\Theta_1\varepsilon_{t-14}\end{aligned} \tag{3.9}$$

The approximate maximum likelihood estimates of the coefficients are

$$\emptyset_1 = 0.6546, \theta_1 = -1.3691\,, \theta_2 = 0.3706, \Theta_1 = -0.9785$$

Thus, the forecasting model for the monthly atmospheric temperature of Saudi Arabia is given by

$$\begin{aligned}\hat{x}_t &= 1.6546x_{t-1} - 0.6546x_{t-2} + x_{t-12} - 1.6546\,x_{t-13} \\ &+0.6546x_{t-14} - 1.3691\,\varepsilon_{t-1} + 0.3706\,\varepsilon_{t-2} - 0.9785\,\varepsilon_{t-12} \\ &+1.3396\,\varepsilon_{t-13} - 0.3626\,\varepsilon_{t-14}\end{aligned} \tag{3.10}$$

To examine the quality of our proposed model, first we graph the forecasting values obtained by our proposed ARIMA $(1,1,2)(0,1,1)_{12}$ model on the top of the original time series data as shown in Figure 3.6 below. As we can see in the plot, the predicted values follow the actual data of the monthly temperature of Saudi Arabia and that an indication of good quality of our proposed forecasting model.

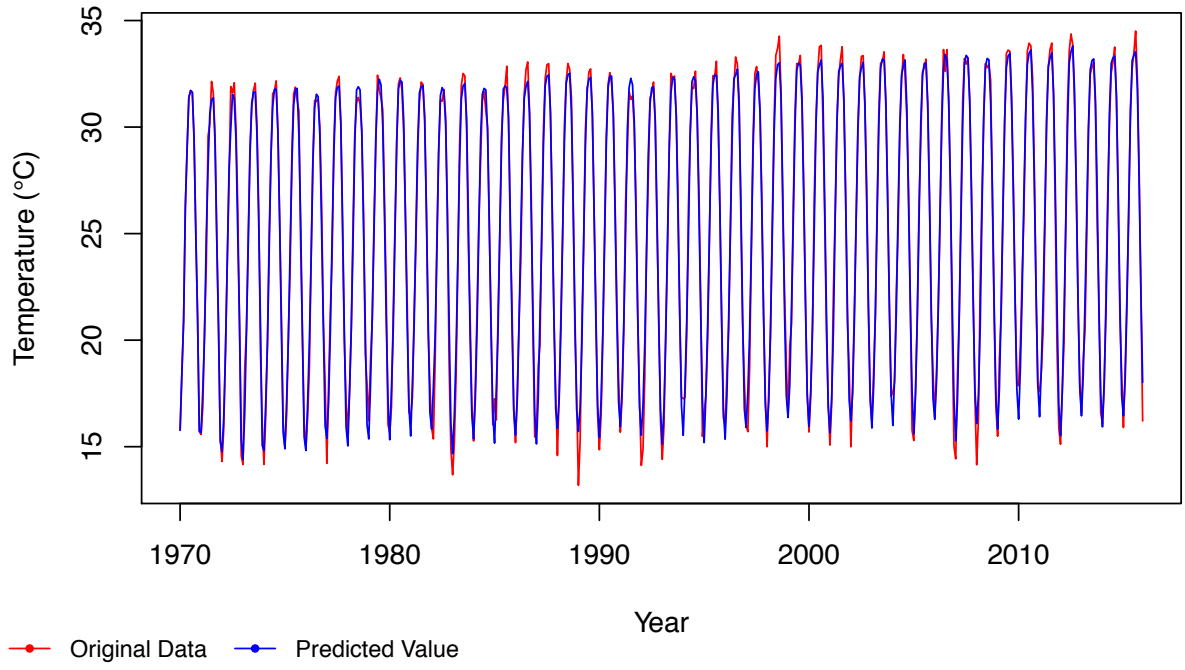


Figure 3. 6 Original vs. predicted values of monthly temperature

Next, we calculate the residuals estimate and evaluate the mean of the residuals, $\bar{r}$, the variance, S_r^2, and the mean square error, *MSE*. The results are presented in Table 3.3 below; Figure 3.7 shows a graphical

presentation of the residual estimates

Table 3. 3 Basic Evaluation of temperature model

$\bar{r}$	S_r^2	*MSE*
0.0451	0.5366	0.5376

The mean of the residuals is very close to zero and it illustrates the best quality of the model, in addition, the residual plot in Figure 3.7 shows that the residual estimated of our proposed model are very small and isolating around zero and the variation of the residuals stays much the same across the time series data. These results also support the effectiveness of the proposed model for forecasting average monthly atmospheric temperature in Saudi Arabia.

Moreover, we restructure model (3.10) again using portion of the data for fitting and use the rest of the data for testing the model. The testing data can be used to measure how well the model is likely to forecast on new data. Table 3.4 gives the 24 hidden values of average monthly temperature, predicted values, and the residuals

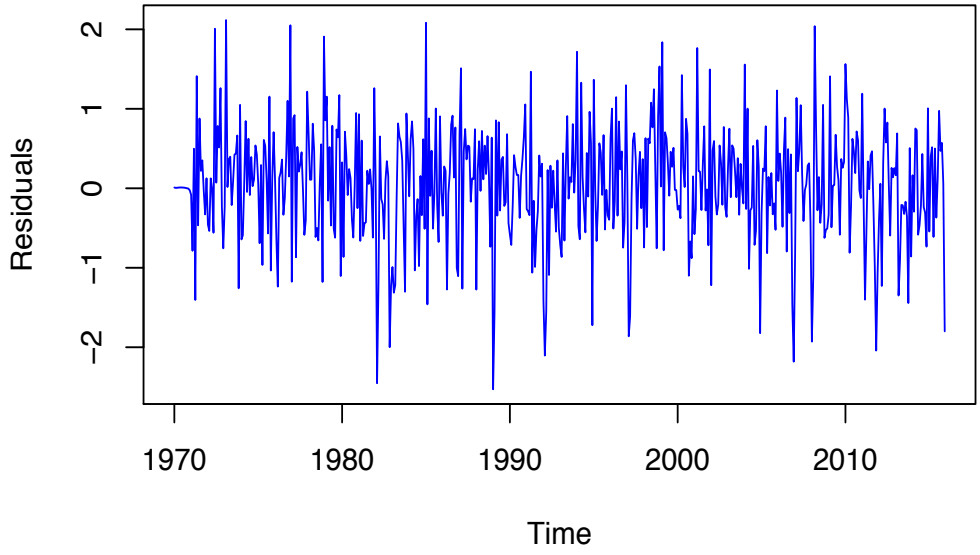


Figure 3. 7 Residual plot of monthly temperature of Saudi Arabia

Table 3. 4 Original data vs. forecasting values of average temperature

Months	Original Values	Forecasting values	Residuals
Jan 2014	16.092	15.91702	0.17498
Feb 2014	17.7728	18.01089	-0.23809
Mar 2104	22.0042	21.28802	0.71618
Apr 2104	26.802	25.87299	0.92901
May 2104	30.1104	30.35522	-0.24482
Jun 2104	32.5334	32.94024	-0.40684
Jul 2104	33.0339	33.34407	-0.31017
Aug 2104	33.7543	33.44256	0.31174
Sep 2104	31.2507	31.43814	-0.18744
Oct 2014	26.689	27.03107	-0.34207
Nov 2014	21.0057	21.84871	-0.84301
Dec 2104	18.3979	17.66596	0.73194
Jan 2015	15.9007	16.33336	-0.43266
Feb 2105	18.6084	18.30048	0.30792
Mar 2015	22.0948	21.49318	0.60162
Apr 2015	25.6254	26.0219	-0.3965
May 2015	30.9111	30.46668	0.44442
Jun 2015	32.7563	33.02674	-0.27044
Jul 2015	33.5017	33.41394	0.08776
Aug 2015	34.5059	33.50136	1.00454
Sep 2105	32.2554	31.48957	0.76583
Oct 2015	27.9791	27.07759	0.90151
Nov 2015	22.3221	21.89195	0.43015
Dec 2105	16.2065	17.70703	-1.50053

The average of these residuals is $\bar{r} = 0.0931$, and Figure 3.8 below shows a graphical result of the predicted values of the average monthly temperature using our proposed forecasting model. Notice how well the forecasts follow the trend in the original data of the average atmospheric temperature in Saudi Arabia, and that is another evidence of the good quality of our proposed forecasting model.

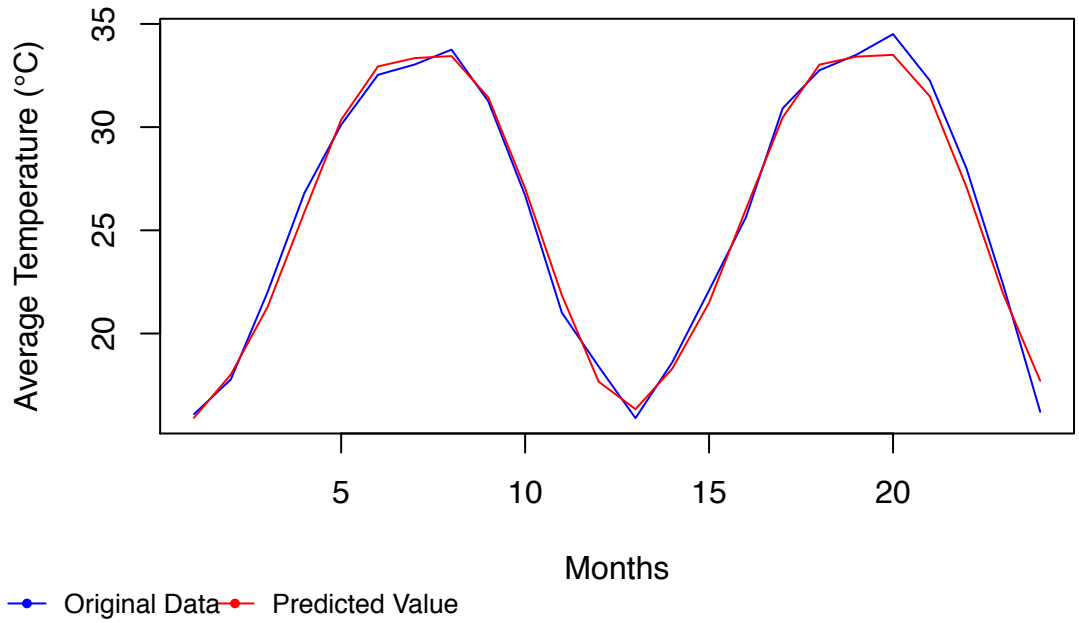


Figure 3. 8 Original data vs. forecasting values of the average temperature

3.4 Conclusion and Contributions

In the present study, we have developed two seasonal autoregressive integrated moving average models to forecast the monthly atmospheric carbon dioxide concentration in the Middle East and the monthly average atmospheric temperature in Saudi Arabia. The two developed statistical forecasting models were evaluated using different statistical criteria; also tested the accuracy of the predicted values and it was shown that both statistical forecasting models produced good estimates. The two forecasting models will help monitor the carbon dioxide emission in the Middle East to the acceptable production amount.

In this study, we were able to accomplish the following goals,

1. We have developed statistical forecasting models of the monthly atmospheric carbon dioxide in the Middle East using the multiplicative seasonal autoregressive integrated moving average model.

2. We have developed a seasonal autoregressive integrated moving average model of the atmospheric temperature in Saudi Arabia.
3. The two forecasting models will help monitor the carbon dioxide emission in the Middle East.
4. The forecasting models assist in developing a strategic policy to maintain the maximum allowable production of carbon dioxide in the Middle East.

Chapter Four

Alzheimer's Disease: The Relative Importance Diagnostic

4.1 Introduction

Alzheimer's disease causes memory loss, and it is not a normal part of aging. It is the only disease that cannot be prevented, treated or even slowed. A recent fact from Alzheimer's Association report in 2018 shows that only deaths from Alzheimer's disease have increased significantly while from other major causes of death in the United States have decreased significantly. The bar chart in Figure 4.1 shows the percentage changes in the top causes of death between 2000 and 2015. As we can see, the number of deaths from heart disease, the number one cause of death in the United States, decreased by 11%; however, recorded death from Alzheimer's disease increased by 123% [24].

In comparison to cancer, 90% of patients become aware of their diagnosis, but only 45 % of the people with Alzheimer's are aware [25]. Thus, researchers and doctors are working to develop a diagnosis pattern of Alzheimer's disease that helps in early detection of the disease before symptoms increase. Different types of tests include neuropsychological test, blood tests, cerebrospinal fluid analysis, and brain imaging have been used to help understand and diagnosis this severe disease. Neuropsychological tests are an assessment of the brain function to evaluate numbers of areas including attention, problem-solving, memory, language, mood, and behavior. Commonly used test tools include the Mini-Mental Status Examination (MMSE) and Dementia Rating Scale (CDR).

Brain imaging is used to detect some brain changes caused by Alzheimer's disease, that is, detecting the levels of plaques and tangles, the two types of disorders in the brain associated with the presence of Alzheimer's. Plaques are found between the dying cells in the brain from the buildup of a protein called

beta-amyloid and tangles are twisted fibers within the dying cells from the other protein called tau. Beta-Amyloid and tau proteins are normally fragmented that the body produces, but in Alzheimer's the proteins are abnormal.

Cerebrospinal fluid analysis (CSF) is collecting the clear fluid that protects and surrounds the brain and spinal cord to determine the levels of beta-amyloid, total tau (T-tau) and phosphorylated tau(P-tau) proteins. Since CSF is in direct contact with the brain and spine, so collecting a sample of the fluid can be a useful diagnostic tool for this neurodegenerative disease.

The primary goal of the present study is to develop the best statistical model to correctly predict Alzheimer's patients with their demographic, CSF, Laboratory and brain imaging factors using logistic regression model. This model will allow us to accurately evaluate the probability that a patient is diagnosed with Alzheimer's disease. Moreover, we can rank the significant contributing risk factors based on their relative importance to the response. Hence, Medical doctor can use our proposed data-driven model as a decision supportive before starting any treatment.

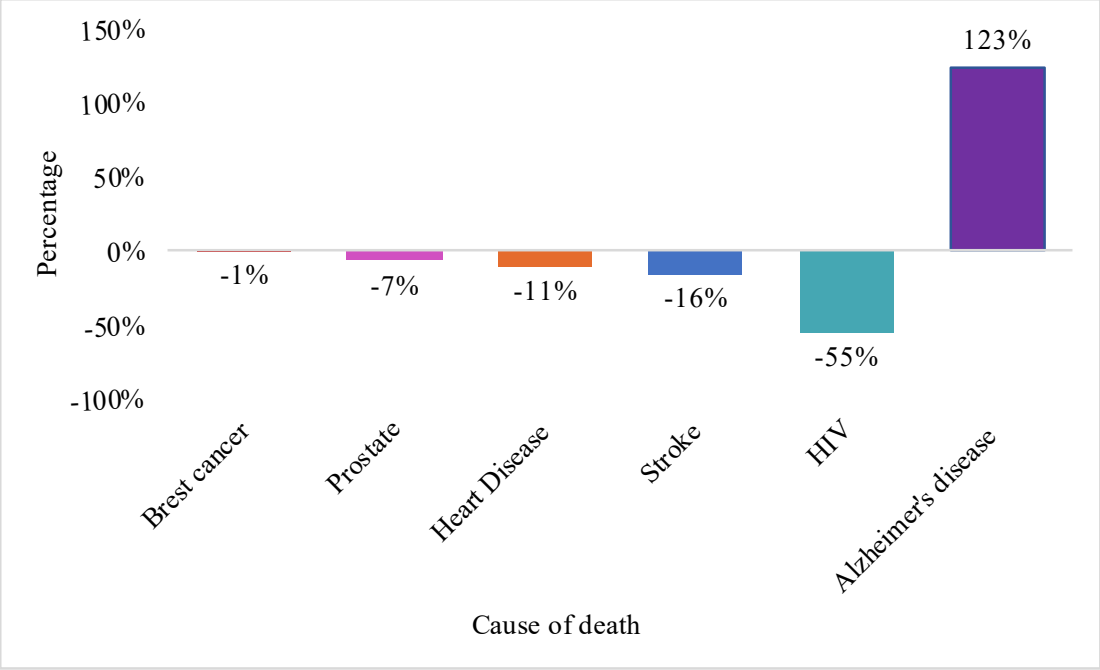


Figure 4. 1 Percentage of selected causes of death in the US between 2000-2015

Source: 2018 Alzheimer's Disease Facts and Figures

4.2 The Data

In the present study, we used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The primary goal of ADNI is to detect and track the progression of Alzheimer's disease by combining clinical, imaging, genetic and biological markers of participants to help researchers and doctors develop new treatments. More information about ADNI visits http://adni.loni.usc.edu.

Our data consist of 169 subjects with an age range from 58-94 years old. We have information about their demographic characteristics, neuropsychological test, laboratory data, cerebrospinal fluid analysis, and brain imaging data. Figure 4.2 below gives an extended detail of our data.

In the cerebrospinal fluid analysis, we have a concentration of P-tau and amyloid beta levels in picograms per milliliter (pg/ml) from the cerebrospinal fluid. The laboratory data consist of the levels of vitamin B12 in nanograms per milliliter (ng/mL), thyroid stimulating hormone in milliunits per liter (mU/L), Hemoglobin in grams per deciliter(g/dL) and cholesterol in milligram per deciliter (mg/dL) as they have been linked to Alzheimer's disease.

MRI scan includes measures about total brain volume, whole brain gray matter volume, whole brain white matter volume, and intracranial volume.

Our response in this Analysis is the status of the participants as cognitively normal individuals (CN) or Alzheimer's disease (AD) based on SPARE-AD score (Spatial Pattern of Abnormalities for Recognition of Early AD). SPARE-AD is an imaging analysis of the spatial patterns of brain atrophy to distinguish individuals with AD from CN. Positive diagnostics values indicate the presence of Alzheimer's disease and negative values indicate a normal pattern of brain structure [26]–[28].

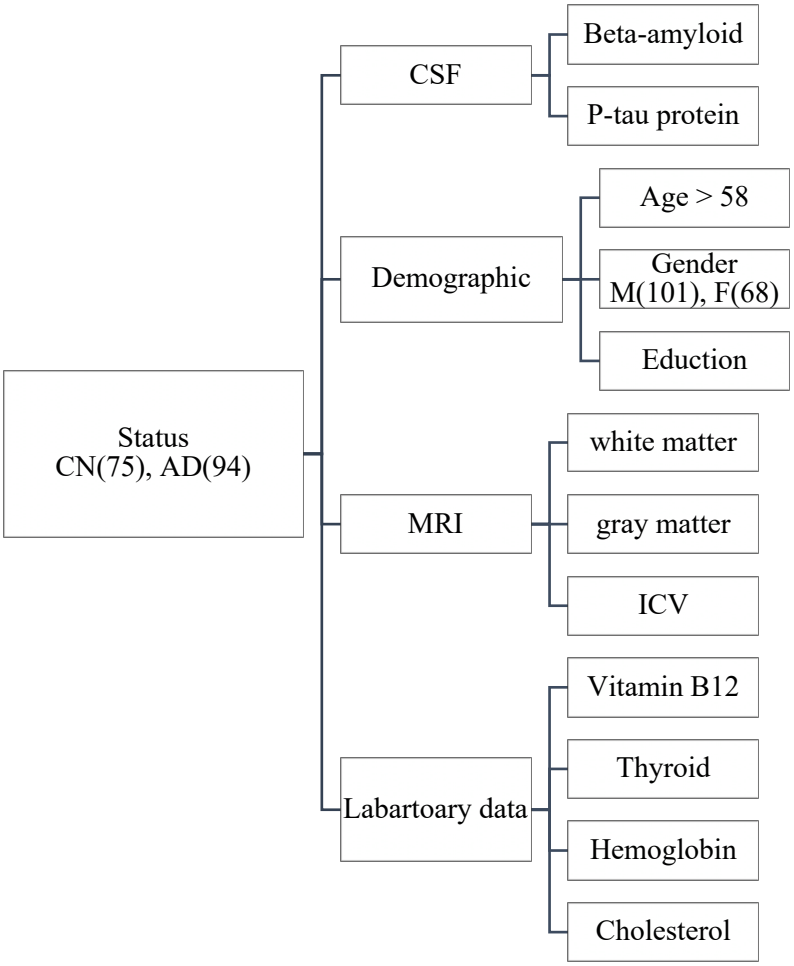


Figure 4. 2 Schematic diagram of the data

4.2.1 Comparison of the probability of Male and Female diagnosed with AD

Several studies have mentioned that women are more likely than men to be identified with Alzheimer's disease [29]. We proceed to investigate this issue by addressing the following question:

- Are male and female equality diagnosed with Alzheimer's disease?

To answer this question, we used the hypothesis test to determine whether the difference between the two proportions is significant. That is, to test the hypothesis that $H_o: P_1 = P_2$ vs. $H_o: P_1 = P_2$, where P_1 is the proportion of male with AD and P_2 is the proportion of female with AD. A *p-value* = 0.7951

indicate that at 5% level of significance, there is no statistically significant difference between the percentage of males and females diagnosed with Alzheimer's disease.

4.3 Statistical Method

For our analysis, we used multiple logistic regression to predict the status of the patients as CN or AD. The logistic regression is a method used to describe and explain the relationship between binary response and the statistically significant risk factors. It can answer questions like: do age, body weight, vitamin B12, cholesterol level, tau, and beta-amyloid proteins influence on the probability of having Alzheimer's disease?

Mathematically, let Y be the binary response and its possible outcome by 1("AD") and 0 ("CN"). The distribution of Y is specified by probability $P(Y = 1) = \pi$ of AD and $P(Y = 0) = (1 - \pi)$ of CN, where $E(Y) = \pi$ is the mean of Y. Let $\pi(x)$ denote the probability of selecting AD patient given the risk factors *x.* The logistic regression model has a linear form for the logit of this probability defined as [30]

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \sum \beta_j x_{ij}\,, \tag{4. 1}$$

where β_j is the coefficient of the j^{th} risk factor $(j = 1, \dots, p)$, x_{ij} is the i^{th} observed value of the risk factor j $(i = 1, \dots, n)$ and$\left(\frac{\pi(x)}{1-\pi(x)}\right)$ is the odds which expresses the ratio between the probability of predicting AD patient to the probability of CN. The logistic regression model implies the analytic for the probability of selecting AD patient given by the risk factors as

$$\pi(x) = \frac{\exp\left(\sum \beta_j x_{ij}\right)}{1 + \exp\left(\sum \beta_j x_{ij}\right)}. \tag{4. 2}$$

4.4 Implementation of the Multiple Logistic Model

We partition our data set into two parts training and testing with 75% and 25% of the data,

respectively. We started with the full logistic regression model that include all predictors and their possible interactions. Our logistic model with all independent variables and their possible interactions to predict whether the patient has Alzheimer's disease is given by

$$\log\left[\frac{P}{1-P}\right] = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j \,, \tag{4. 3}$$

Where P denote the probability of selecting AD patient, $\beta_j's$ denote the coefficients and $X's$ are the risk factors and possible interactions. Using backward elimination algorithm to remove the term in the complex model that has the largest P_value and stop when any further elimination leads to poor fit. In addition to the minimum AIC (*Akaike information criterion*) that judges the quality of the model by how close the fitted values to the true expected values, that means, selecting the best statistical predictive model that minimize

$$AIC = -2\ln(L) + 2k,$$

where L is the value of the likelihood and k is the number of parameters in the model. Thus, our optimal data-driven statistical logistic model that predicts the patient's condition with minimum AIC is given by:

$$\begin{aligned}\log\left[\frac{P}{1-P}\right] &= 7.55 - 0.003\,Abeta + 0.170\,PTau + 10.18\,Thyroid \\ &+0.002\,VB12 - 0.14Chelost - 0.44\,Hemog + 0.01\,(Chelost \cap Hemog) \\ &- 0.87\,(Thyroid \cap Hemog)\end{aligned} \tag{4. 4}$$

The symbol (∩) means interaction and as we can see from our proposed model, only six risk factors and two interaction terms are statistically significant contributing to the prediction of the patient's condition, namely, phosphorylated tau protein (P-tau), beta-amyloid protein, thyroid stimulating hormone, vitamin B12, cholesterol, hemoglobin, and the interaction between (cholesterol ∩ hemoglobin) and (thyroid stimulating hormone ∩ hemoglobin). Furthermore, as we can see, age is not one of the significant risk factors in our optimal predictive model, and this holds that Alzheimer's disease is not part of normal aging.

The coefficients in the logistic regression indicate the change in the expected log odds relative to the one-unit change in (X_j) holding all other predictors are constant [31], [32]. Thus, the interpretation of

the coefficient (0.170) of P-tau protein means as the P-tau protein level increases, the odds of the participant diagnosed with AD will increases while holding all other variables constant. Alternatively, we can use the odds ratio $exp\ (0.170) = 1.85$, and that means with all other predictors unchanged, every unit increase in the P-tau protein increase the odds of beaning Alzheimer's patient by a factor of 1.85.

Similarly, the interpretation of the coefficient (-0.003) of beta-amyloid protein means that as the beta-amyloid protein level decrease, the odds of the participant diagnosed with AD will increase while holding all other variables constant. Alternatively, by using the odds ration $exp\ (-0.003) = 0.997$, with all other predictors unchanged, every unit decrease in the beta-amyloid protein increases the odds of being Alzheimer's patient by a factor of 0.997.

4.4.1 Model Evaluation

To evaluate our optimal predictive model, we used classification accuracy, sensitivity, specificity values and area under the curve (AUC) for testing data. The proportions of correctly identified AD and CN participants from the multiple logistic model is called "accuracy". The proportions of actual Alzheimer's patients who are correctly identified from our predictive model as having the disease is known as "sensitivity" and the proportions of actual cognitively normal individuals who are correctly identified from the model is known as "specificity". A perfect predictive model would be described as 100% sensitive (that is predicting all sick people from Alzheimer's disease group as Alzheimer's) and 100% specific (that is predicting all normal individual as cognitively normal). For any test, however, there is usually a trade-off between these two measures and can be explored graphically by the receiver operating characteristic curve (ROC).

We used the confusion matrix of the testing data to get the values needed to assess the model. The confusion matrix is a classification table describe how well our multiple logistic regression model does in predicting Alzheimer's patients from cognitively normal individuals. Table 4.1 shows an illustration of a confusion matrix that we used to evaluate our proposed model on the test data. The four outcomes that

formulated the table are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP is the number of Alzheimer's patients correctly identified as sick, and TN is the number of normal individuals correctly classified as healthy. FP is the number of healthy people incorrectly identified as sick, and FN is the number of Alzheimer's cases predicted incorrectly by our model as a healthy individual.

Using the confusion matrix, we found out that our model accuracy is $\left(\frac{TP+TN}{N+P}\right) = 80\%$ and it correctly predicts 78.26% of all Alzheimer's disease cases (the sensitivity = $\left(\frac{TP}{P}\right)$). Also, it correctly identifies 83.33% of those who don't have Alzheimer's disease (the specificity=$\left(\frac{TN}{N}\right)$). A summary of our classification results is given in Table 4.2 below.

Another method to evaluate our model graphically is the receiver operating characteristic (ROC). Each point on the ROC curve represents a (sensitivity,1-specificity) pair corresponding to a different decision cut-off point. The area under the ROC curve (AUC) is a measure of how well the model can distinguish between two diagnostic groups. For our proposed model, the AUC value is 87.68% which implies that our model does well in discriminating between the two classes of the patient's condition. Figure 4.3 represents the receiver operating characteristic curve with the corresponding AUC value. After a careful investigation of our results, we can conclude that our predictive model provides a good prediction of the patient's condition.

Table 4. 1 The confusion matrix

		Actual class		**Total**
Predicted class		CN	AD	
	CN	TN = 10	FN = 5	15
	AD	FP = 2	TP = 18	20
Total		N =12	P = 23	35

Table 4. 2 Classification summary of the multiple logistic regression model

Evaluation value	Percentage
Accuracy	80%
Sensitivity	78.26%
Specificity	83.33%

After validating our proposed model, we need to rank the risk factors in terms of their importance to Alzheimer's diagnostic. We identified the relative importance of the risk factors by the absolute value of their standardized coefficients (weights) and pseudo partial correlation. In the standardized coefficients, the higher the absolute value points to the greater strength of association with Alzheimer's diagnostic [33], [34]. The standardized weight is defined as

$$\text{Standardized weight} = \frac{\beta_i}{(s/sd_i)}, \tag{4. 5}$$

where β_i is the estimated coefficient (weight) for predictor i, sd_i is the sample standard deviation for predictor i, and $s = \pi/\sqrt{3}$.

The pseudo partial correlation is given by

$$r = \pm\sqrt{(W_i - 2K)/-2LL_0} \tag{4. 6}$$

where W_i is the Wald chi-square statistic for predictor i, K is the degrees of freedom of predictor i, and $-2LL_0$ is the log-likelihood of the model with only intercept term. The closer the value to 1 or -1, the stronger the association between a predictor and the outcome, [35].

Thus, the relative importance of the significantly contributing risk factors in our predictive model is presented in Table 4.1. As can be seen, the result of the two methods is consistent, and we found out that P-tau protein is the most critical factor in diagnosing with Alzheimer's disease followed by beta-amyloid. These two proteins have been extensively studied by the author [36]. Also, the interaction between (thyroid ∩ hemoglobin) is ranked as number three significant predictor before the level of thyroid hormone alone and hemoglobin alone which they ranked as number 4th and number 8th significant risk factors, respectively.

Table 4. 3 Relative importance of the risk factors

Rank	Risk Factor	Standardized Weights	Pseudo Partial Correlation
1	P-Tau protein	4.384	0.542
2	Beta-amyloid	3.568	-0.410
3	Thyroid ∩ Hemoglobin	2.514	-0.243
4	Thyroid	2.171	0.212
5	Vitamin B12	1.665	0.196
6	Cholesterol	1.554	-0.154
7	Cholesterol ∩ Hemoglobin	1.496	0.147
8	Hemoglobin	0.349	-0.019

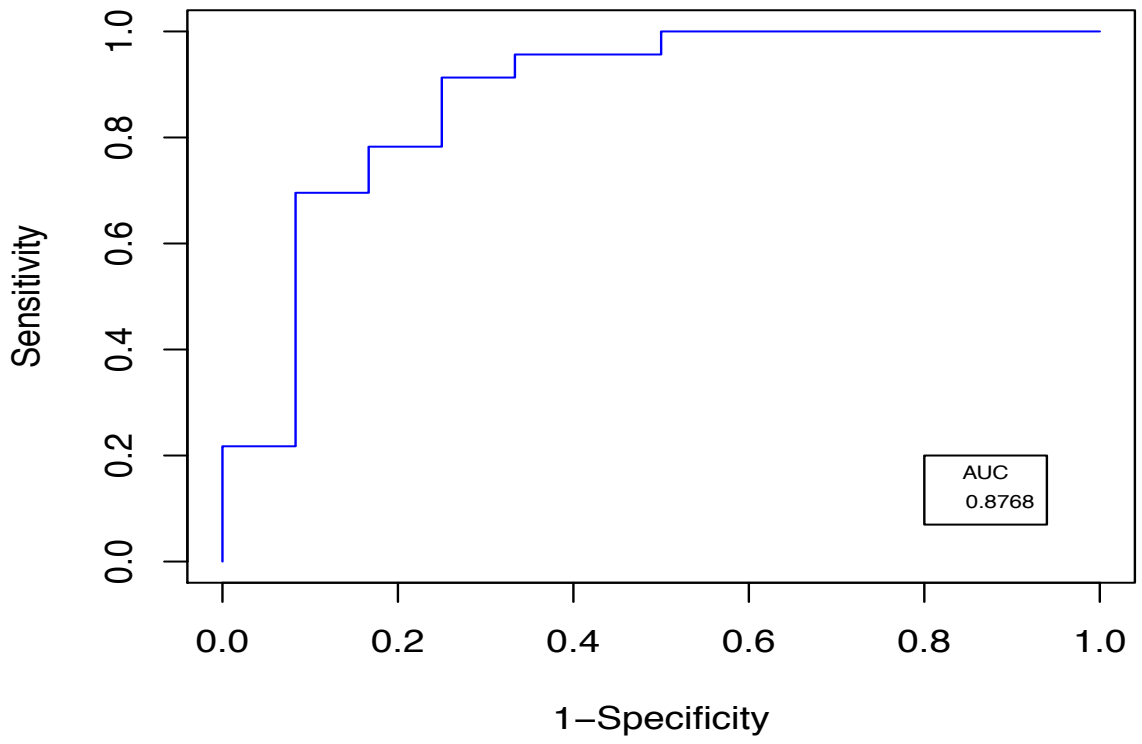


Figure 4. 3 The receiver operating characteristic curve

4.5 Conclusion and Contributions

The importance of knowing the causes of the disease helps find the best way to cure it. While

several top causes of death are decreasing, Alzheimer's deaths are on the rise. Thus, in the present study, we developed a statistical predictive model using multiple logistic regression to predict Alzheimer's disease patients by selecting the relevant risk factors using backward elimination. We found that only six risk factors and two interaction terms namely, phosphorylated tau protein (P-tau), beta-amyloid protein, thyroid stimulating hormone, vitamin B12, cholesterol, hemoglobin, and the interaction between (cholesterol ∩ hemoglobin) and (thyroid stimulating hormone ∩ hemoglobin) are significantly contributing to Alzheimer's disease.

We evaluated the quality of the proposed model by classification accuracy, sensitivity, specificity values and area under the curve, the result of which attest to the effectiveness of the model. Then, we examine the relationship between the response and the significant contributing predictors and rank them based on their standardized coefficients. By defining and ranking of the statistically significant risk factors, they will be useful as a screening tool to discriminate Alzheimer's disease patients from cognitively normal individuals.

In this study, we were able to accomplish the following goals,

1. We show that at 5% level of significance, there is no statistically significant difference between the proportion of males and females diagnosed with Alzheimer's disease as several studies mentioned that women are more likely than men to be identified with Alzheimer's disease.
2. We have developed an effective diagnosis statistical predictive model using multiple logistic regression to predict Alzheimer's disease.
3. The proposed analytics multiple logistic regression model can identify the relevant risk factors of Alzheimer's disease and proceed for medical treatments if necessary.
4. Age is not one of the significant risk factors in our optimal predictive model, and this holds that Alzheimer's disease is not part of normal aging.
5. The information obtained from our proposed statistical model would avoid unnecessary treatments and improve the financial aspects.

Chapter Five

Alzheimer's: A Probabilistic Approach to the Behavior of Beta-Amyloid and Phosphorylated Tau Proteins Level

5.1 Introduction

Alzheimer's disease is the most common form of dementia and a serious disease affecting the brain. It is an invisible disease that destroys memory and important mental functions slowly so that patient in their final stage of life cannot assume the simplest daily tasks. The patient's actions become irrational and lose the ability to think and control their behavior. Alzheimer's disease is not a normal stage of aging, but the possibility of infection increases with age. According to the Alzheimer's Association, an estimated 5.5 million Americans of 65 years of age and older have Alzheimer's disease in 2017 and it is the 6^{th} leading cause of death in the United States that cannot be prevented, cured or even slowed.

The cerebrospinal fluid (CSF) is a clear and colorless liquid that surrounds the brain and spinal cord to protect the central nervous system. Since CSF is in direct contact with the brain, biochemical changes in the brain are reflected in the CSF. Two abnormal structures called plaques and tangles are prime suspects in damaging the nerve cells in the brain with Alzheimer's. Plaques are found between the dying cells in the brain from the buildup of beta-amyloid protein (β) and tangles are twisted fibers of phosphorylated tau protein $(P\tau)$ that buildup inside the cells. Figure 5.1 below shows the difference between the size of a healthy brain and Alzheimer's disease, and Figure 5.2 shows the healthy brain nerve cells and cell destroyed by plaques and tangles in Alzheimer's disease. Beta-Amyloid and $P\tau$ proteins are normally fragmented that the body produces, but in Alzheimer's the proteins are abnormal. Thus, the combinations

of increased cerebrospinal fluid levels of $P\tau$ protein and decreased level of beta-amyloid have been suggested as possible diagnostic contributors to Alzheimer's disease (AD) [37]–[39].

In the present study, our data was obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI data is a non-treatment study data from multiple centers across the United States and Canada and their primary goal is to examine and analyze the progression of Alzheimer's disease from combined biological, clinical, brain imaging and neuropsychological assessments data. For more information on ANDI data, visit http://adni.loni.usc.edu. A total of 210 records of measured $P\tau$ (min=9.89, max=60.73) and beta amyloid (min=212.3, max=1664) levels in pg/ml from the cerebrospinal fluid of subjects participated in ADNI study were used to perform the statistical parametric analysis. Figure 5.3 gives an extended detail of our data. All subjects entered into the ADNI database underwent a blood test, cerebrospinal fluid analysis, brain imaging and standardized behavioral assessment such as Mini Mental State Examination (MMSE) and the Clinical Dementia Rating (CDR), which they are used to measure dementia severity of Alzheimer's patients. MMSE is considered to be effective as a screening tool and one of the most commonly used rating scales. The subject procedure tests five areas of cognitive function namely, orientation, registration, recall, attention and calculation [40].

In this study, we use CSF levels (pg/ml) of $P\tau$ and beta-amyloid to identify the probability distribution function (pdf) that probabilistically characterizes their behavior separately and the maximum likelihood estimates of the parameters along with appropriate degree of confidence. In addition, to understand the probabilistic abnormality behavior of beta-amyloid and tau proteins happening at the same point of time by constructing their bivariate probability distribution function using the copula method.

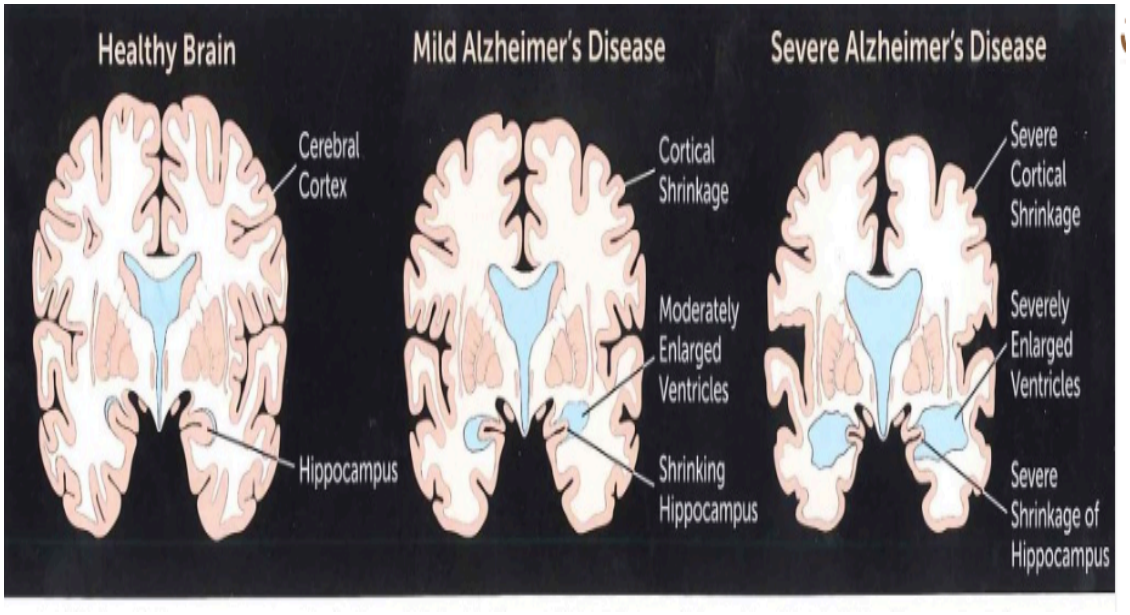
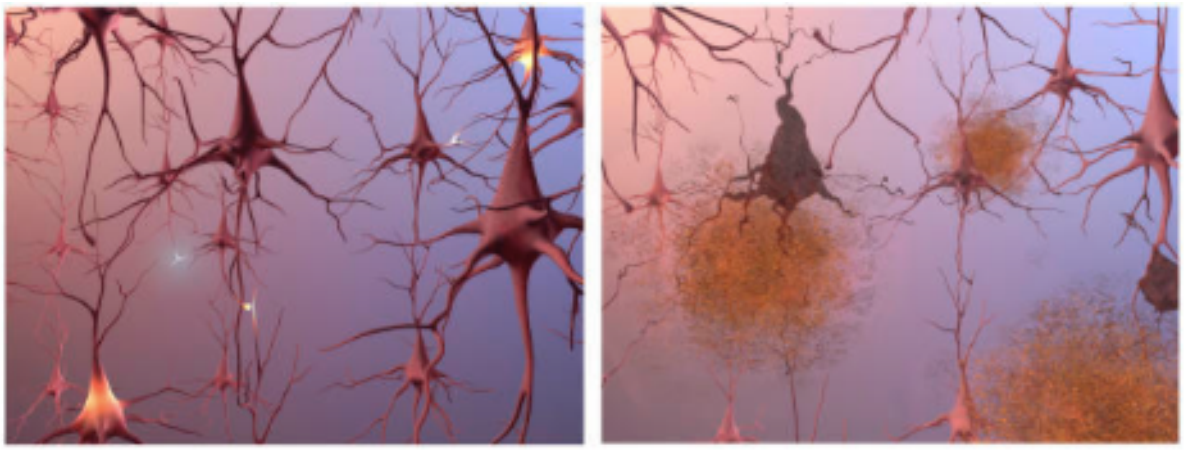


Figure 5. 1 Healthy brain vs. Alzheimer's disease

*Source: http://memorylanecottage.com/about-alzheimers-and-dementia/how-the-brain-changes-during-alzheimers-disease/



(a) (b)

Figure 5. 2 (a) Healthy brain cells and (b) Alzheimer's disease cells with plaques and tangles

*Source: Alzheimer's Association: https://www.alz.org/alzheimers-dementia/what-is-alzheimers/brain_tour

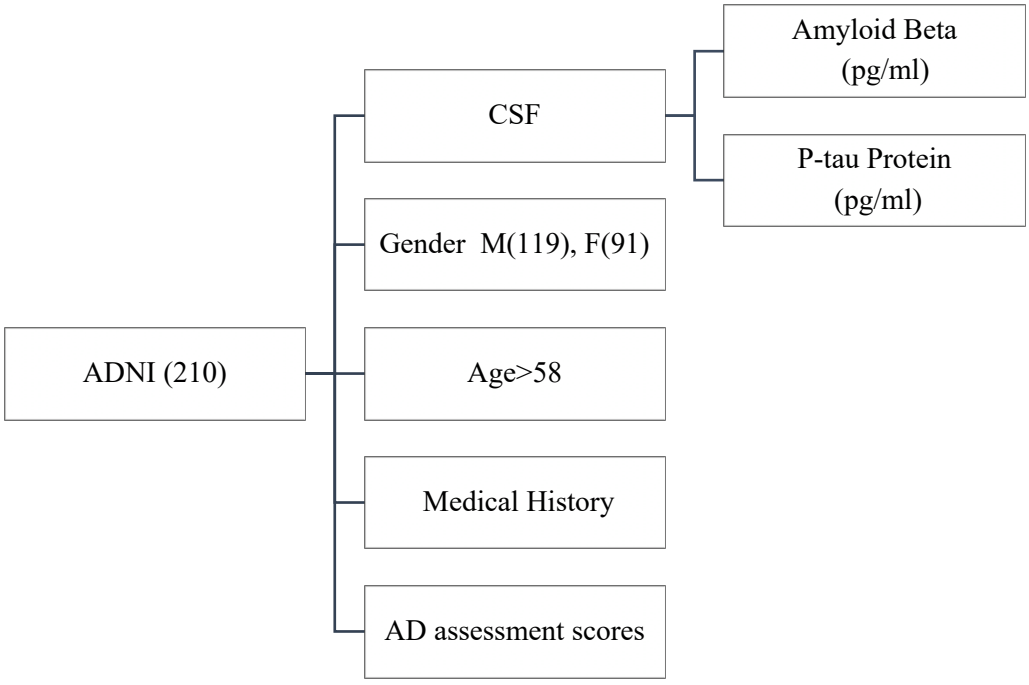


Figure 5. 3 Schematic diagram of Alzheimer's data

5.2 Copula

The behavior of a single variable is fully described by its probability distribution. In multiple variables of interest, all the information on the behavior is fully described by the joint probability distribution which measures the likelihood of two events occurring together at the same point of time. Recently, copula becomes a popular method in constructing a bivariate and multivariate distributions with pre-defined marginal distributions and correlation their coefficient. Different fields discovered the importance of this method for building more flexible multivariate distributions that can take any probability distribution, which does not need to be equal for all the margins. Unlike the multivariate normal probability distribution requires its marginals to be normally distributed.

All the methods and theory of copula hold for the multivariate case, thus, for our study, we only consider the two-dimensional case. A two- dimensional copula is a function $C(.,.)$: $[0,1]^2 \rightarrow [0,1]$, that is,

a function to link two-dimensional probability distributions to their one-dimensional margins satisfying the following properties:

Let $u = F_{x_1}(x_1), v = F_{x_2}(x_2)$ denoting the marginal probability distributions of X_1 and X_2, respectively, then the properties are [41]:

1. C is grounded, i.e., for every $(u, v) \in [0,1]^2, C(u, 0) = C(0, v) = 0$. This property means that if the realization of one variable has the marginal probability zero, then the joint probability of all outcomes is zero.
2. $C(u, 1) = u \; and \; C(1, v) = v$ for every $(u, v) \in [0,1]^2$. This property implies that if the realization of one of the variables is known with marginal probability one, then the joint probability is equal to the one with uncertain outcome.
3. C is two- increasing.

Copula is a Latin word which means "link, couple" and was first introduced by Sklar (1959) who obtained and proved the most important result in this area by introducing Sklar's theorem. That is,

- **Sklar's Theorem:** Let $F(x_1, x_2)$ be the joint cumulative probability distribution function with the marginal cumulative probability distributions $F_{x_1}(x_1), F_{x_2}(x_2)$, there exists a copula C, such that

$$F(x_1, x_2) = C\left(F_{x_1}(x_1), F_{x_2}(x_2)\right) \tag{5. 1}$$

for all $x_i \in [-\infty, \infty]$, $i = 1,2$. This theorem states that the joint cumulative probability distribution function can be written as of marginal cumulative probability distribution functions and copula, which describes the dependence between the variables. Conversely, if C is bivariate copula and $F_{x_1}(x_1), F_{x_2}(x_2)$ are the cumulative probability distribution functions, then the function $F(x_1, x_2)$ defined in (5.1) is a bivariate probability distribution function with cumulative marginals $F_{x_1}(x_1), F_{x_2}(x_2)$. Therefore, the probability density function of the bivariate probability distribution can be written as

$$f(x_1, x_2) = c\left(F_{x_1}(x_1), F_{x_2}(x_2)\right) f_1(x_1), f_2(x_2), \tag{5. 2}$$

where $f_{x_1}(x_1), f_{x_2}(x_2)$ are the probability density functions corresponding to the cumulative probability distribution functions $F_{x_1}(x_1), F_{x_2}(x_2)$ and $c\left(F_{x_1}(x_1), F_{x_2}(x_2)\right) = \frac{\partial^2 C(F_1(x), F_2(x))}{\partial F_1(x) \partial F_2(x)}$ is the copula density [42]. There are several copula functions that can be used to construct the bivariate probability distribution with given marginals, the selection process depends on the strength of the correlation coefficient.

5.2.1 Classes of Copulas:

In this section, we present the most commonly used copula functions and their properties. There are two parametric classes of copulas: implicit and explicit copulas. Implicit copulas are derived by well-known multivariate probability distributions. The most known copulas from this family are Gaussian copula and t- copula. The Gaussian copula is implied by multivariate normal distribution and multivariate Student t-distribution leads to t-copula.

- **Gaussian Copula:** The bivariate Gaussian copula is given by

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} exp\left\{-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho 2)}\right\} dx_1 dx_2 \qquad (5.3)$$

where the linear correlation coefficient $\rho \in [-1,1]$ is the dependence parameter of the copula and Φ is the inverse univariate standard normal distribution.

- **Student t-copula:** The bivariate t-copula is given by

$$C_t(u, v) = \int_{-\infty}^{t^{-1}(u)} \int_{-\infty}^{t^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \left(1 + \frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{v(1-\rho 2)}\right) dx_1 dx_2, \qquad (5.4)$$

with dependence parameter ρ and degree of freedom v for the student t-copula, while the t^{-1} is the inverse univariate Student-t distribution function with v degree of freedom, expected value 0 and variance $\frac{v}{v-2}$.

Explicit copulas are also called Archimedean copulas which are not derived from multivariate distribution functions but do have simple closed forms. We will consider two Archimedean copulas: Clayton and Gumbel copulas[43], [44].

- **Clayton copula:** Clayton has defined the copula by

$$C_{Cl}(u,v) = \left(u^{-\delta} + v^{-\delta} - 1\right)^{-1/\delta}, \tag{5. 5}$$

where $\delta \epsilon (0, \infty)$ is the copula parameter, if $\delta \to \infty$ implies dependence while $\delta \to 0$ implies independence.

- **Gumbel copula:** The bivariate Gumbel copula takes the following form

$$C_{Gu}(u,v) = exp\left(\left[(-\log u)^{\delta} + (-\log v)^{\delta}\right]^{1/\delta}\right), \tag{5. 6}$$

where $\delta \geq 1$ is the dependence parameter of the copula and when $\delta \to \infty$,we have a perfect dependence and when $\delta = 1$ implies independence between the two variables.

In addition to these families, there are rotated versions of copulas. The 90- and 270-degrees rotated copulas allow for the modeling of negative dependence, which is not possible to model with regular copula, while rotating them by 180 degree, we get the corresponding survival copula. The distribution functions of the rotated copula *C* by 90, 180 and 270 degrees, respectively, are given by

$$C_{90}(u,v) = v - C(1-u,v), \tag{5. 7}$$

$$C_{180}(u,v) = u + v - 1 + C(1-u,1-v), \tag{5. 8}$$

and

$$C_{270}(u,v) = u - C(u,1-v). \tag{5. 9}$$

5.2.2 Process of Selecting the Copula:

The appropriate copula function that provides the best fit to the given data can be selected by comparing the evaluated values of the Akaike Information Criterion, AIC, which is defined as

$$AIC = -2\ln(L) + 2k, \tag{5. 10}$$

where *L* is the value of the likelihood and *k* is the number of parameters of the copula model. The copula associated with the smallest AIC value provides the best fit [45], [46].

5.3 Result

5.3.1 Comparison of Mean CSF levels of Phosphorylated Tau and Amyloid Beta between Genders

Several studies have mentioned that women are more likely, than men, to be identified with Alzheimer's dieses [29]. We proceed to investigate this issue by addressing following question:

- Is there a significant difference in the mean levels of amyloid beta and $P\tau$ protein between males and females of Alzheimer's patients?

To answer this question, we perform non-parametric analysis (assumptions free) test: The Kruskal-Wallis test to determine whether there are statistically significant differences between mean levels of P-tau and beta-amyloid proteins between genders, respectively. For $P\tau$ level difference, using the following notations, μ_1, μ_2 to represent the true population mean of $P\tau$ protein level for males and females, respectively, we use the Kruskal-Wallis to test whether the two population means are equal or not, i.e. $H_o: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$. The samples mean of CSF level of $P\tau$ protein and sample standard deviations for males are 27.68 pg/ml and 11.80, respectively, while for females are 28.51 pg/ml and 11.79, respectively. A p-value of 0.552 indicate that at 5% level of significant, there is no statistically significant difference in CSF level of $P\tau$ protein between males and females.

Similarly, for CSF level of beta-amyloid protein, the sample mean and sample standard deviation for male is 746.64 pg/ml and 388.03, respectively, while for female is 822.07 pg/ml and 395.45, respectively. Let μ_M and μ_F symbolize the true mean of CSF level of amyloid beta for male and female, respectively. For testing $H_o: \mu_M = \mu_F$, we have concluded with a p-value=0.1475 that there is no statistically significant difference between the true means of beta-amyloid protein level between males and females. Thus, we can combine the data of males and females to perform our analysis.

5.3.2 Parametric Analysis

We proceed to perform parametric analysis of CSF of $P\tau$ and amyloid beta protein levels data, respectively. Statistically we shall first show that the data does not follow the commonly used classical Gaussian probability distribution, which leads to misunderstanding the behavior of the subject data. Secondly, we shall identify the best-fit probability distribution function (pdf) that characterizes the levels of $P\tau$ and beta-amyloid proteins, respectively.

The phosphorylated tau level has a sample mean of 28.04 pg/ml with a sample standard deviation of 11.77 and beta-amyloid level has a sample mean of 779.33 pg/ml and a sample standard deviation of 392.11. Both proteins level data does not display symmetry nor bell-shaped smoothness as shown in the histogram in Figure 5.4 and Figure 5.5, respectively.

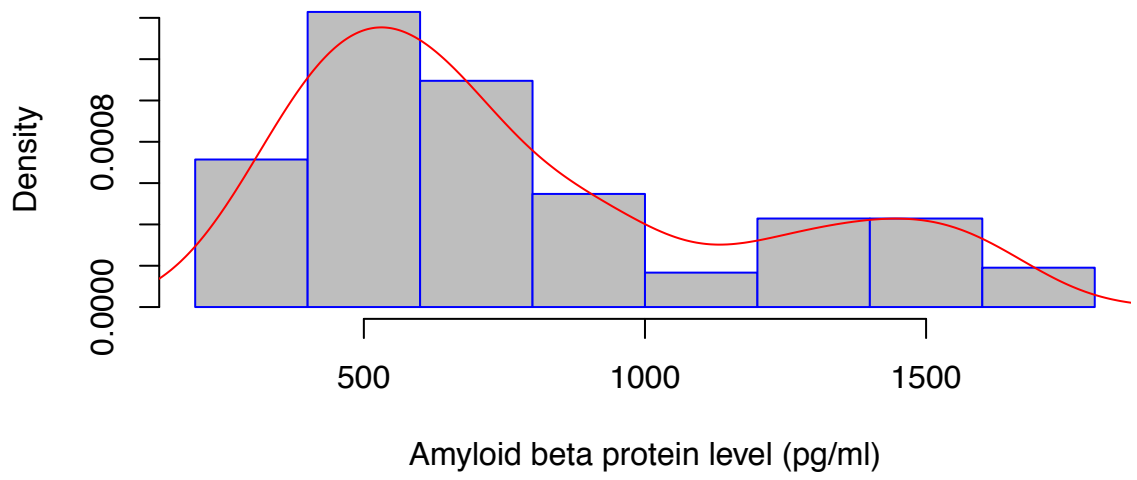


Figure 5. 4 Histogram of beta-amyloid

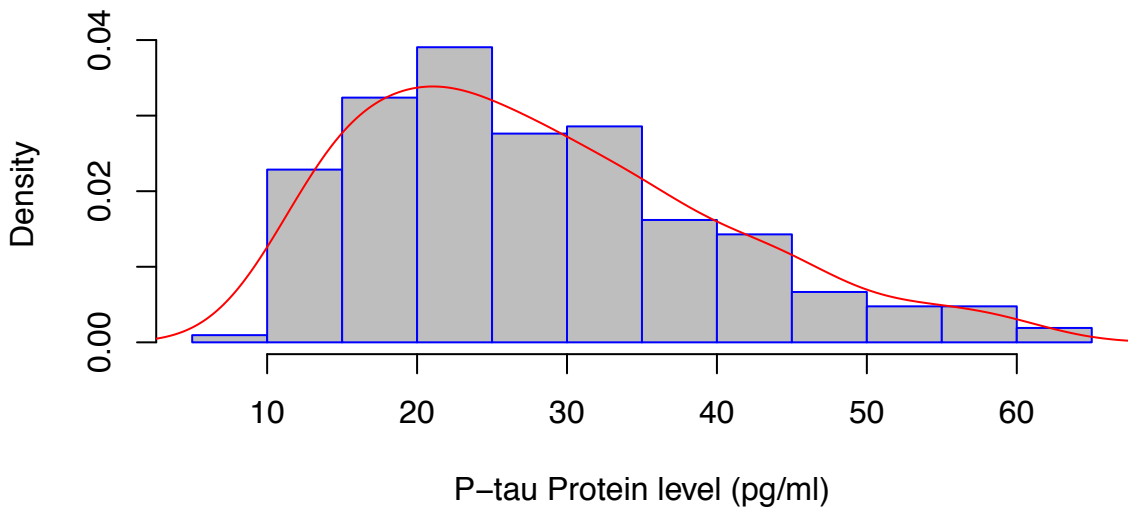


Figure 5. 5 Histogram of P-tau protein

For the best possible probability distribution that characterizes the behavior of $P\tau$ and beta-amyloid proteins, we tested different probability distributions using three standard statistical tests: Kolomogrov-Smirnov [8], Anderson-Darling [9] and Chi-square [10]. The Kolmogorov-Smirnov test is based on minimum difference estimation. Anderson-Darling measures whether the data can transform into the uniform probability distribution and the Chi-square test for goodness of fit is a measure of relative error squared. Using these three tests, the Gaussian pdf did not pass any of the three tests.

5.3.2.1 Probability Distribution Function of Phosphorylated Tau Level

The three-parameter Weibull probability distribution best characterizes the probabilistic behavior of the $P\tau$, and the pdf is given by

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right), \qquad \gamma \leq x \tag{5. 11}$$

Where $\alpha > 0$ is the shape parameter, $\beta > 0$ is a scale parameter and γ is a location parameter. The corresponding cumulative distribution function is given by

$$F(x) = 1 - exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right). \quad (5.\ 12)$$

The approximate maximum likelihood estimates of the parameters α, β, and γ are given in Table 5.1.

Table 5. 1 Approximate maximum likelihood parameters estimate, expected value and standard deviation

Parameters	Approximate estimate
$\widehat{\boldsymbol{\alpha}}$	1.588
$\widehat{\boldsymbol{\beta}}$	20.636
$\widehat{\boldsymbol{\gamma}}$	9.492
Expected value	28.01
Standard deviation	11.928

With respect to the probability distribution, the expected level of P-tau protein of a randomly selected patient is going to be 28.01 pg/ml.

Thus, the pdf of the three-parameter Weibull probability distribution with $9.492 \leq x$ is given by

$$f(x) = 0.013(-9.492 + x)^{0.588}\, exp(-0.008(-9.492 + x)^{1.588}), \quad (5.\ 13)$$

and the corresponding CDF of the three-parameter Weibull probability distribution is given by

$$F(x) = 1 - exp(-0.008(-9.492 + x)^{1.588})\,. \quad (5.\ 14)$$

The graph of the pdf and CDF of the $P\tau$ protein is given below by Figure 5.6 and Figure 5.7, respectively. Thus, finding the probability of protein level of randomly selected patient involves finding the area under the curve as shown in the graph of the pdf of $P\tau$ protein in Figure 5.6 below. That is, the probability of randomly selected patient will have $P\tau$ protein level between 40 pg/ml and 65 pg/ml i.e. $P(40 \leq X \leq 65) \approx 0.1475$.

We can use the cumulative probability distribution function (CDF) to obtain any estimates of the cumulative probability of a random observation taken from the population will be less than, exceed a certain critical level value that is desirable or between two values. That is, using CDF, we can calculate the cumulative probability that a randomly chosen patient from such a population will be less than a certain level of $P\tau$ protein. For instance, as shown in Figure 5.7 below, $P(X < 28)$= 0.5688, is the cumulative

probability of $P\tau$ protein level of a randomly chosen patient from such a population will be below the average.

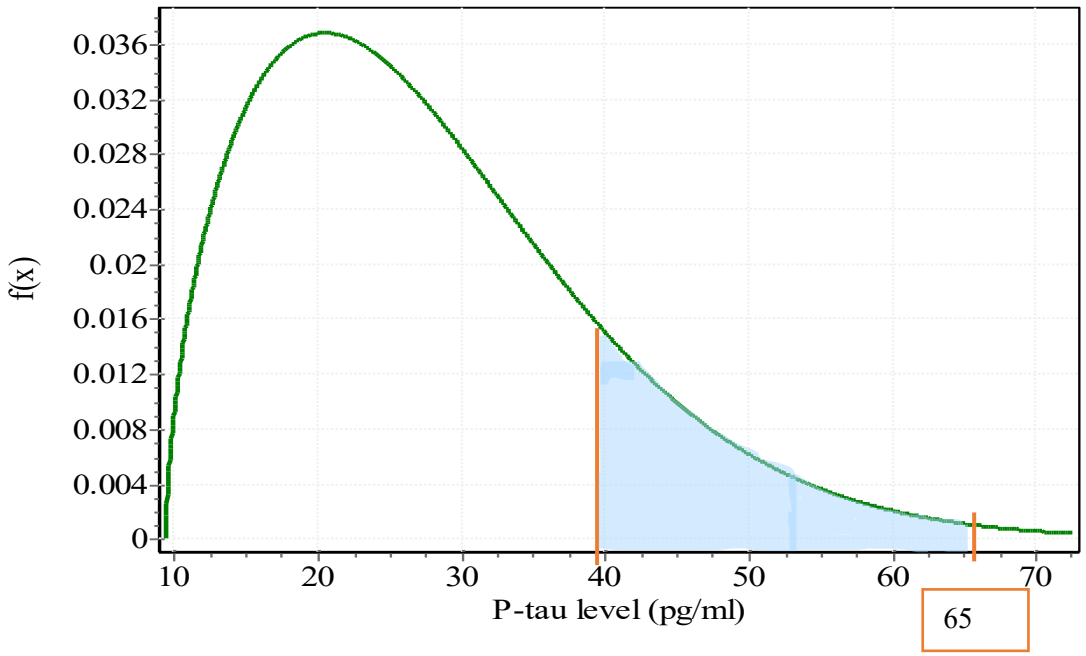


Figure 5. 6 Probability distribution function plot of P-Tau protein level

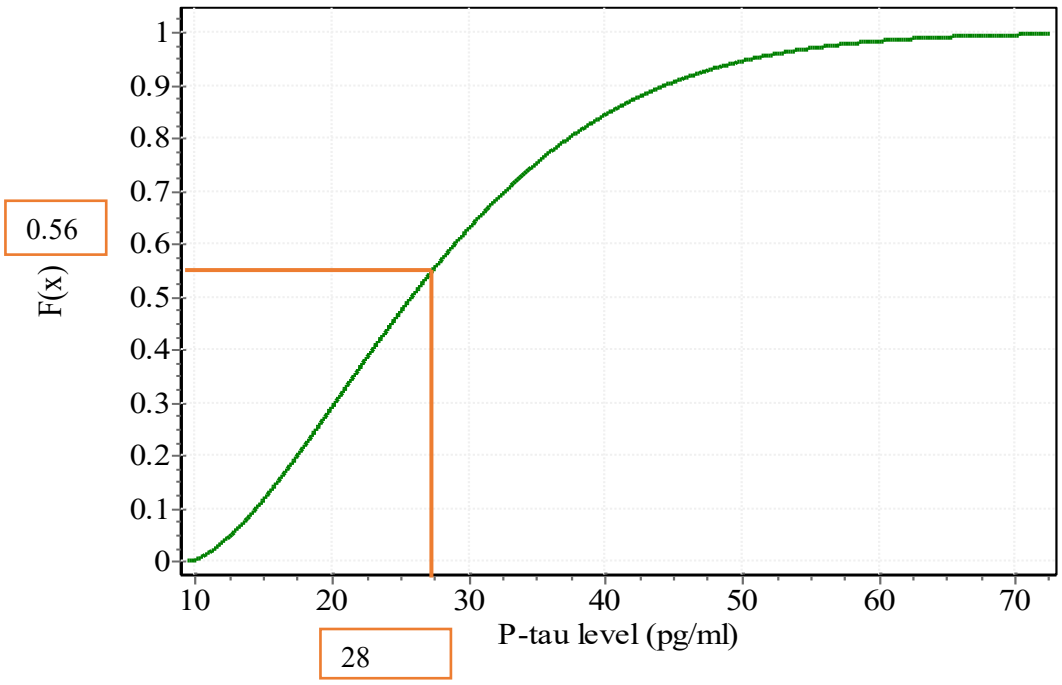


Figure 5. 7 Cumulative distribution function plot of P-Tau protein level

5.3.2.2 Probability Distribution Function of Beta-Amyloid Level

The three parameters log-logistic distribution is the best probability distribution that fits or explains the beta-amyloid level data and its pdf is given by

$$f(y) = \frac{\alpha}{\beta}\left(\frac{y-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{y-\gamma}{\beta}\right)^{\alpha}\right)^{-2}; \gamma \leq y < \infty, \quad (5.\ 15)$$

where $\alpha > 0$ is the shape parameter, $\beta > 0$ is a scale parameter and $\gamma > 0$ is a location parameter. The corresponding cumulative distribution function is given by

$$F(y) = \left(1+\left(\frac{\beta}{y-\gamma}\right)^{\alpha}\right)^{-1}. \quad (5.\ 16)$$

The approximate Maximum Likelihood Estimates (MLE) of the parameters α, β and γ, in Table 5.2. With respect to the three parameters log-logistic distribution, the expected level of amyloid beta protein of a randomly selected patient is 823.54 pg/ml.

Table 5. 2 Approximate MLE of the parameters

Parameters	MLE
$\widehat{\boldsymbol{\alpha}}$	2.511
$\widehat{\boldsymbol{\beta}}$	498.13
$\widehat{\boldsymbol{\gamma}}$	167.12
Expected value	823.54
Standard deviation	780.43

Thus, the pdf of three parameters log-logistic distribution is given by

$$f(y) = \frac{6.574 * 10^{-6}(-498.13+y)^{1.511}}{(1+2.618 * 10^{-6}(-498.13+y)^{2.511})^2}; \quad 167.12 \leq y, \quad (5.\ 17)$$

and the corresponding CDF of the three parameters log-logistic distribution is given by

$$F(y) = \frac{1}{1+381966\left(\frac{1}{-498.13+y}\right)^{2.511}} \quad (5.\ 18)$$

The graphical view of the probability distribution function and cumulative distribution function of the three parameters log-logistic probability distribution that characterize the level of Aβ protein is given by Figure 5.8 and Figure 5.9, respectively. Figure 5.8 shows the probability of randomly chosen patient will have beta-amyloid level between 500 pg/ml and 1000 pg/ml, i.e. $P(500 \leq Y \leq 1000) \approx 0.5178$. Figure 5.9 represents the cumulative probabilities of a patient chosen at random that will have a value less than or equal to certain level of beta-amyloid, for example, $P(Y \leq 779) \approx 0.626$ or it will exceed 779 pg/ml is $P(Y \geq 779) = 1 - P(Y \leq 779) \approx 0.374$.

In addition, we proceed to obtain approximate estimates of at least 90, 95 and 99% confidence limits for the true mean value of $P\tau$ and beta-amyloid proteins separately, that is, finding an interval of the true unknown mean of $P\tau$ protein or beta-amyloid level that an individual selected randomly from such a population will be at least 90%, 95% and 99% certain that it falls between the given intervals. The estimated intervals for $P\tau$ and beta-amyloid protein are shown in Table 5.3 below.

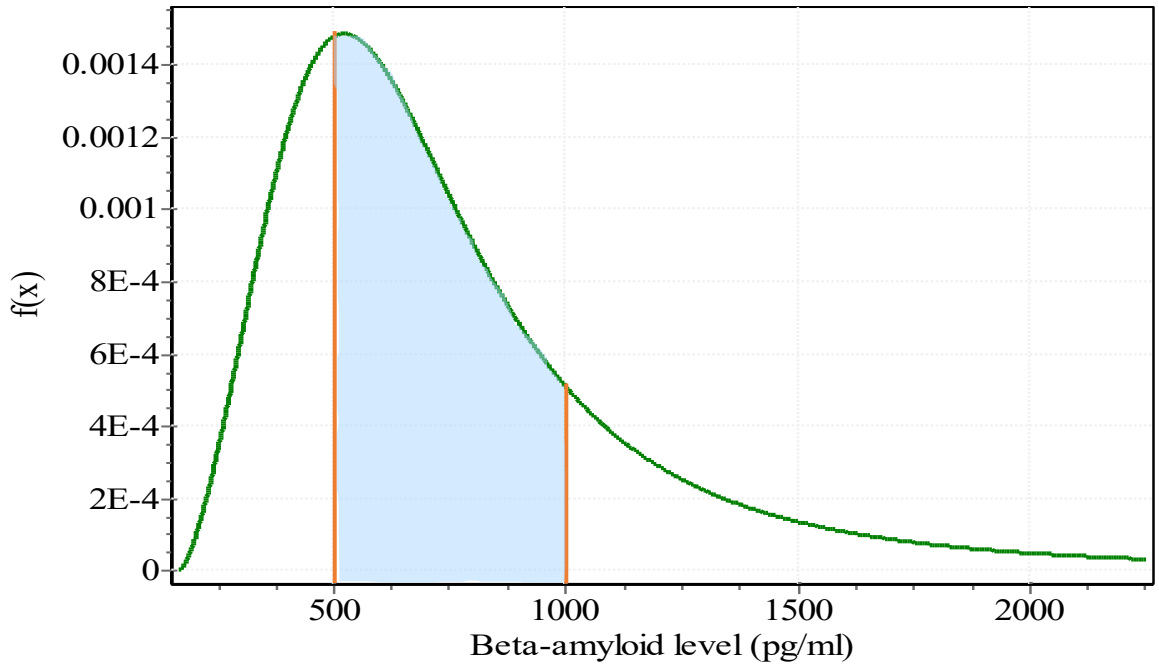


Figure 5. 8 Probability distribution function plot of beta-amyloid level

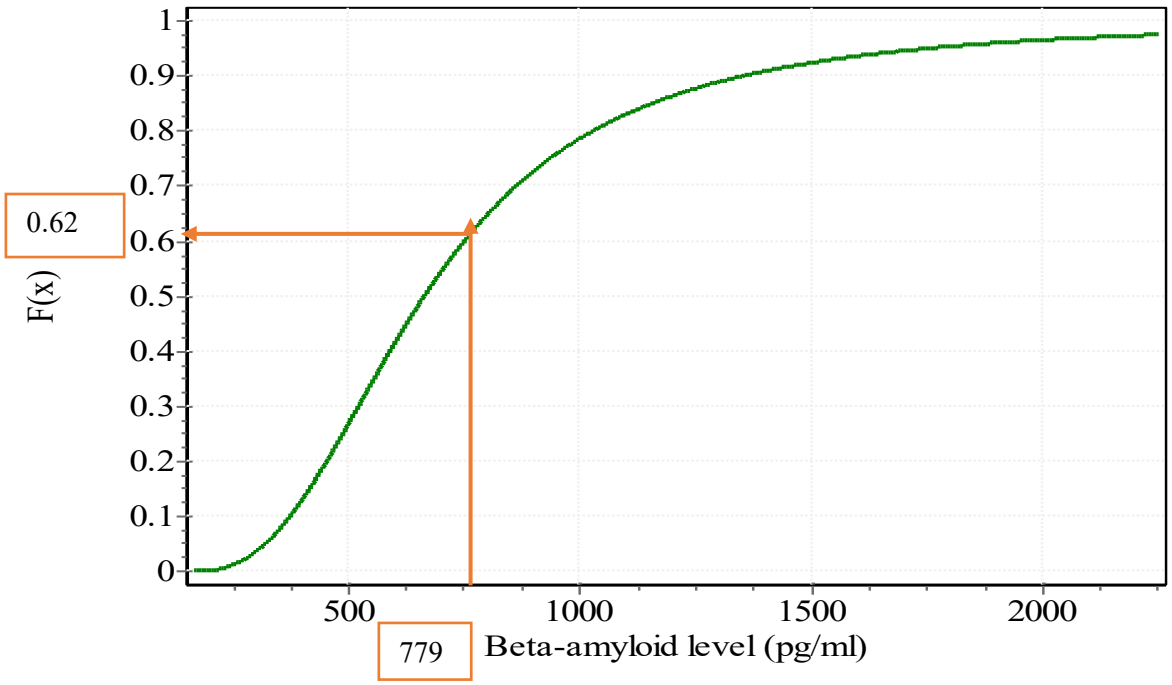


Figure 5. 9 Cumulative distribution function of beta-amyloid level

Table 5. 3 Confidence Limit of the true mean of the two proteins level

Level (%)	CI of $\boldsymbol{P\tau au}$	CR	CI of beta-amyloid	CR
90	(26.65, 29.36)	2.71	(734.94, 912.13)	177.18
95	(26.39, 29.62)	3.22	(717.98, 929.1)	211.12
99	(25.88, 30.13)	4.25	(684.59, 962.48)	277.89

5.3.3 Bivariate Distribution of Beta-Amyloid and Phosphorylated Tau Proteins:

Developing the joint probability distribution function of $P\tau$ and beta-amyloid, we will be able to study the probabilistic behavior of the two proteins as related to Alzheimer's disease. In this section, we will construct a data driven bivariate probability distribution function of $P\tau$ and beta-amyloid proteins using copula function [47]–[50]. Since the correlation between phosphorylated-tau and beta-amyloid is $\hat{\rho} = -0.41$, the best copula function that fit our data with minimum AIC is 90-degree rotated Joe-Frank (BB8) copula. The Joe-Frank copula $C(u, v)$ is given by

$$C(u,v) = \delta^{-1}\left[1 - \left\{1 - \left[1-(1-\delta)^{\theta}\right]^{-1}\left[1-(1-\delta u)^{\theta}\right]\left[1-(1-\delta v)^{\theta}\right]\right\}^{\frac{1}{\theta}}\right], \qquad (5.\ 19)$$

where $\theta \in [1,\infty)$ and $\delta \in (0,1]$. By substituting Equation (17) in Equation (7), we obtain the 90-degree rotated Joe-Frank (BB8) copula as

$$C(u,v) = v - \delta^{-1}\left[1 - \left\{1 - \left[1-(1-\delta)^{\theta}\right]^{-1}\left[1-\left(1-\delta(1-u)\right)^{\theta}\right]\left[1-(1-\delta v)^{\theta}\right]\right\}^{\frac{1}{\theta}}\right], \qquad (5.\ 20)$$

where $\theta \in (-\infty,-1)$ and $\delta[-1,0)$.

the approximate maximum likelihood estimates of the copula parameters are $\theta = -3.55$ and $\delta = -0.55$, then the joint cumulative probability distribution function as in Equation (5.1) can be written as

$$F(x,y) = \left\{\frac{1}{1+381966\left(\frac{1}{-498.13+y}\right)^{2.511}} + 1.82\left(1-\frac{1}{A}\right)\right\}, \qquad (5.\ 21)$$

where

$$\mathrm{A} = \left[1 - 1.267\left[1 - \frac{1}{[1+0.55\,exp(-0.008(-9.49+x)^{1.588})]^{3.55}}\right]\right]\left[1 - \frac{1}{\left[1+\frac{0.55}{1+381966\left(\frac{1}{-498.13+y}\right)^{2.511}}\right]^{3.55}}\right]^{0.28}.$$

Using Equation (5.2), the joint probability density function with the three-parameters Weibull and three parameters log-logistic probability distribution functions driven from 90-degree rotated Joe-Frank (BB8) copula with $x \geq 9.492$ and $y \geq 498.13$ becomes

$$f(x,y) = \frac{8.523 * 10^{-8}\,(-9.492+x)^{0.588}\,(-498.13+y)^{1.511}\,exp(-0.008(-9.492+x)^{1.588})}{(1+2.618*10^{-6}(-498.13+y)^{2.511})^2}\,[A+B], \tag{5. 22}$$

where

$$A = \frac{2.477}{\left[a^{4.55}\left[1-1.267\left[1-\frac{1}{a^{3.55}}\right]\left[1-\frac{1}{b^{3.55}}\right]\right]\right]^{1.282}\,(b)^{4.55}},$$

and

$$B = \frac{\left[4.024\left[1-\frac{1}{a^{3.55}}\right]\left[1-\frac{1}{b^{3.55}}\right]\right]}{a^{4.55}\left[1-1.267\left[1-\frac{1}{a^{3.55}}\right]\left[1-\frac{1}{b^{3.55}}\right]\right]^{2.282}\,(b)^{4.55}},$$

Where

$$a = \left(1+0.55\,exp(-0.008(-9.4923+x)^{1.588})\right)$$

and

$$b = \left[1+\frac{0.55}{1+381966\left(\frac{1}{-948.13+y}\right)^{2.511}}\right].$$

The graphical presentation of the joint probability distribution is presented in Figure 5.10, below. Figure 5.11 gives a rotated graph from different angles to explore possible relationships between the two proteins. As we can see, the peak on the plot occurs at approximately $P\tau \approx 20$ pg/ml and beta-amyloid level≈ 600 pg/ml. That is, the higher probability corresponds to approximately 20 pg/ml and 600 pg/ml of $P\tau$ and beta-amyloid protein level, respectively. Having a bivariate distribution, we can calculate different characterization of the two proteins behavior, that means, the probability when two proteins are in the average levels, or the probability if one within the average level and other one exceeds or below the average level, i.e.,

$$P(x \leq 46, y \leq 600) = \iint f\,(x, y)\,dx\,dy$$

$$= \iint \frac{8.523 * 10^{-8}\,(-9.492 + x)^{0.588}\,(-498.13 + y)^{1.511}\,exp(-0.008(-9.492 + x)^{1.588})}{(1 + 2.618 * 10^{-6}(-498.13 + y)^{2.511})^2}\,[A + B]\,dx\,dy$$

$$\approx 0.000054,$$

that is, the probability of randomly selected patient will have $P\tau$ protein level less than 46 pg/ml and beta-amyloid level less than 600 pg/ml is approximately 0.00005. Thus, understanding the bivariate behavior of $P\tau$ and beta-amyloid can help in finding a drug to control their level as believed that they are suspected signs of Alzheimer's disease.

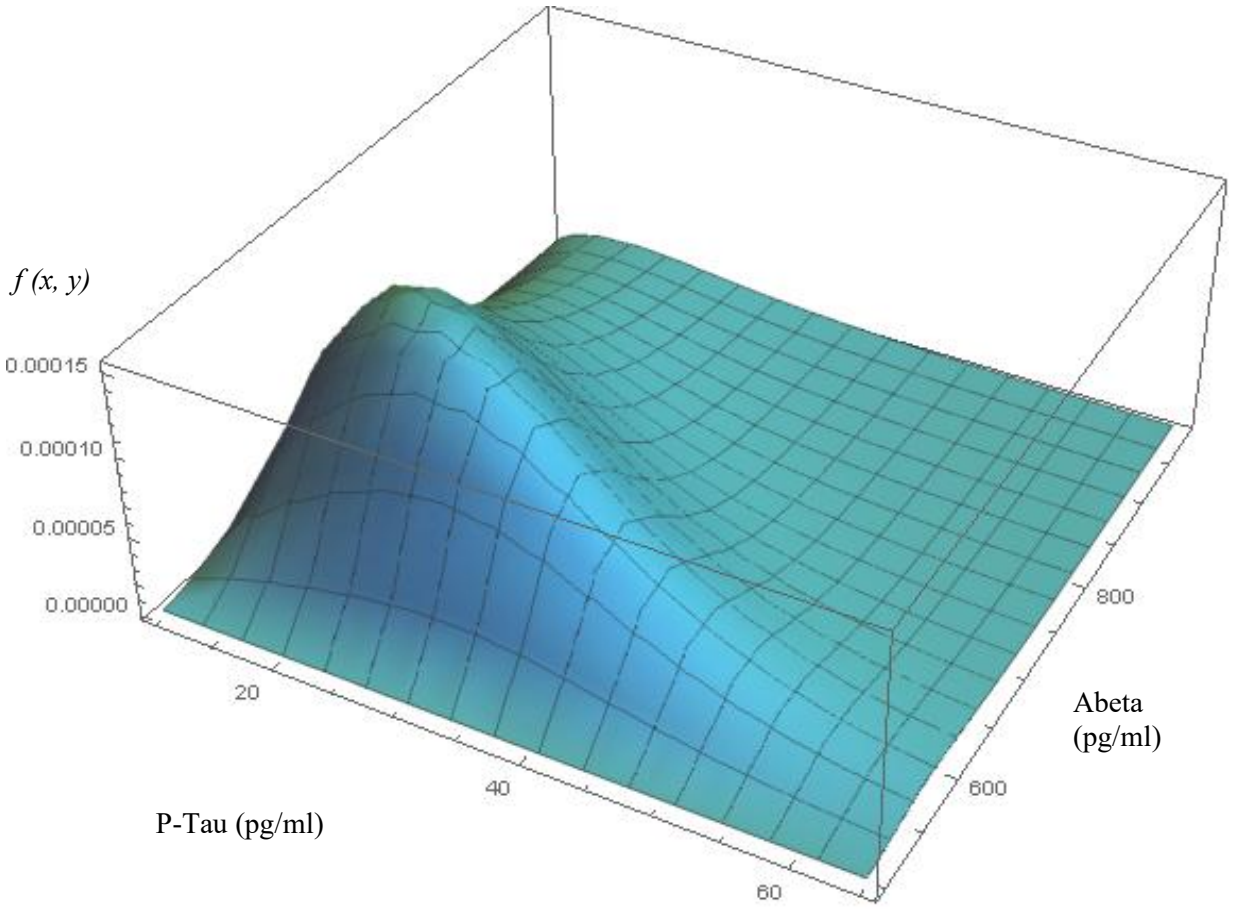


Figure 5. 10 3D plot of the bivariate distribution function of P-tau and beta-amyloid proteins.

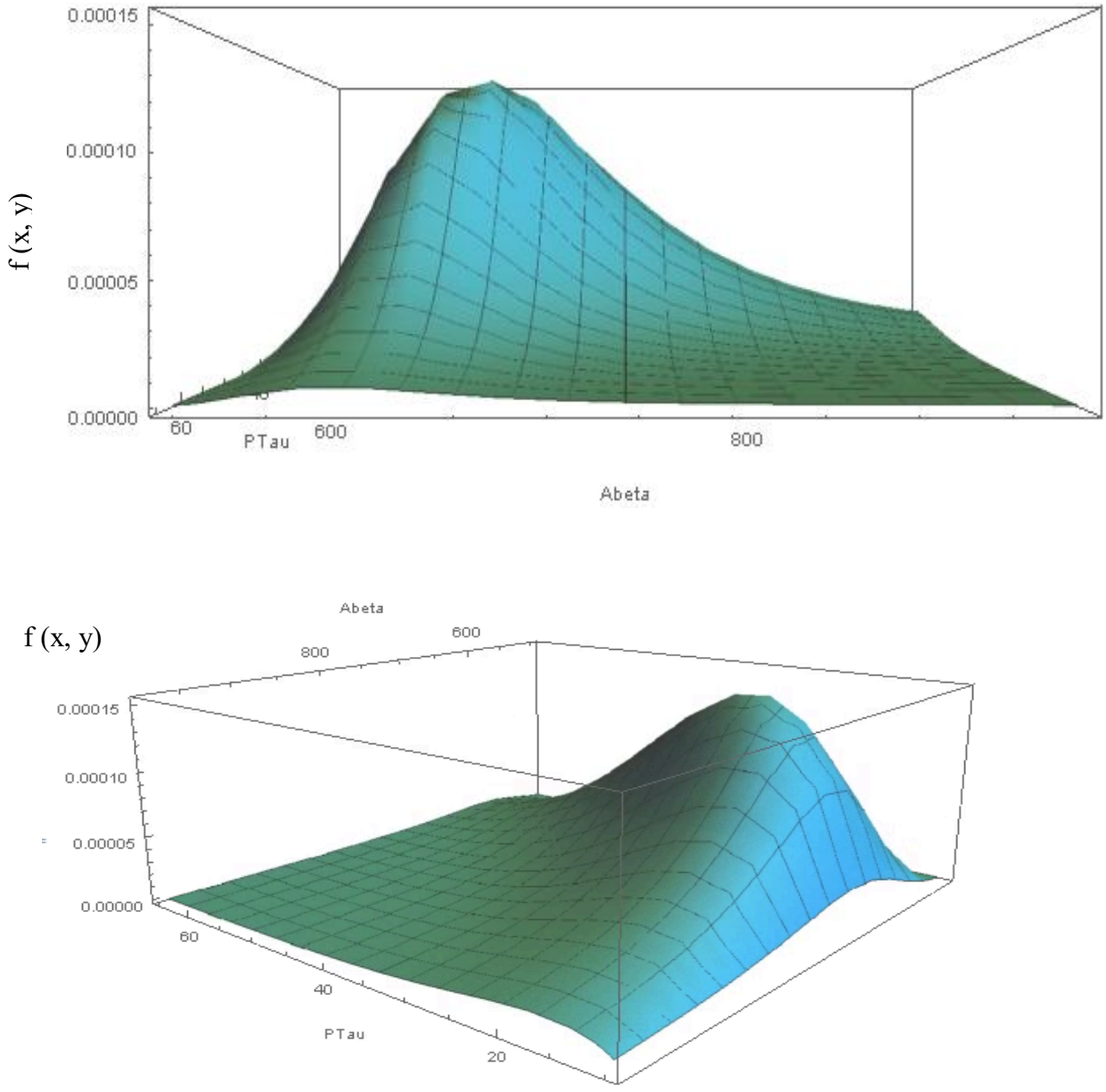


Figure 5. 11 The joint probability distribution plot from different angles

5.4 Justification

Copula parametrically specified joint probability distribution generated from the given marginals, thus the properties of copulas are similar to the properties of joint pdf's that satisfies the following:

- $f(x,y) \geq 0, for\ x,y\ \in \mathbb{R}$.

Since C is two- increasing, this means that the joint probability will not be negative because the $C - value$ of any two-dimensional interval is positive.

- $\int_{167.12}^{\infty} \int_{7.77}^{73.08} f(x, y)\, dx\, dy = 1,$

This has been proven using Mathematica software [51].

5.5 Conclusion and Contributions

In the present study, we performed a parametric analysis of the cerebrospinal fluid (CSF) levels of $P\tau$ protein and beta-amyloid of Alzheimer's patients. We examined the level of $P\tau$ and beta-amyloid proteins between gender, separately, using non-parametric tests to determine if their levels are significantly different to guide us in our study. We have shown that there is no significant difference in the true mean of the level of $P\tau$ and beta-amyloid protein between male and female.

Secondly, we have found that the Gaussian pdf is statistically rejected to probabilistically characterize the behavior of both proteins and have identified that the three-parameters Weibull probability distribution and the three parameters log-logistic probability distribution are the best to describe the probabilistic behavior of $P\tau$ and beta-amyloid proteins, respectively. With the maximum likelihood estimates of the parameters, we calculate the fundamental properties of both proteins level and construct at least 90, 95 and 99% confident intervals of the true mean level under the fitted probability distributions.

Finally, we developed the joint behavior of subject variables of interest $P\tau$ and beta-amyloid levels by constructing their bivariate probability distribution with specific marginals we have identified, and the calculated value of their correlation. We used the copula method that links the marginal probability distributions together to form the joint probability distribution. We found that 90-degree rotated Joe-Frank (BB8) copula function is the best copula function that best fits our data. Thus, obtaining the joint probability distribution of the two proteins, we can calculate different characterization of their bivariate behavior and finding a drug that can control their levels.

Our contributions to this chapter can be summarized as follows:

1. At 5% level of significance, we have shown that there is no significant difference in the true mean of $P\tau$ and beta-amyloid protein level between male and female
2. We have identified that the three-parameters Weibull probability distribution and the three parameters log-logistic probability distribution are the best to characterize the probabilistic behavior of phosphorylated tau and beta-amyloid proteins, respectively.
3. Using copula, we have developed the joint probability distribution of phosphorylated tau and beta-amyloid proteins that characterize their bivariate behavior.
4. Obtaining the joint probability distribution of the two proteins, we can calculate different characterization of their bivariate behavior and finding a drug that can control their levels.

Chapter Six

Future Research

6.1 Regional Analysis of the Atmospheric Carbon dioxide in the Middle East

One of the critical issues on our planet is climate change which is the increase in the atmospheric temperature caused by the rise of the carbon dioxide emission from the industrial revolution. We want to proceed with statistical modeling of carbon dioxide in different regions in Saudi Arabia based on different carbon dioxide emission sectors such as industrial, commercial, electrical, transportation, and residential. The regional probability models will predict the probabilities of the carbon dioxide emission at risk in each region based on values of attributable variables in the previous year then rank the significant risk factors based on their contribution to the response. These models will provide guidelines for the policymaker to control the carbon dioxide emission in each province.

6.2 Mathematical Characterization of Beta-Amyloid and Phosphorylated Tau proteins as a Function of Age

Alzheimer's disease is not a normal part of aging, but the majority of people with Alzheimer's disease are 65 and older. Also, the changing rate of beta-amyloid and phosphorylated tau proteins is the pathological marker of the disease. Thus, the goal of this study is to find the best analytical changing behavior of the two proteins to investigate the effect of patient's age on their levels, separately.

6.3 Statistical Classification Model to Distinguish Alzheimer's disease from Different Types of Dementia

Many tests are being conducted to help distinguish Alzheimer's disease from other memory loss diseases. It can be difficult to discriminate Alzheimer's disease from other dementia since there is an overlap in many common clinical features. For that, it is increasingly important to develop a statistical classification model to diagnose dementia types correctly.

References

[1] National Institute on Aging, "Assessing Cognitive Impairment in Older Patients Benefits of Early Screening," *Https://Www.Nia.Nih.Gov/Alzheimers/Publication/Assessing-Cognitive-Impairment-Older-Patients*, 2014.

[2] M. I. Habadi and C. P. Tsokos, "Statistical Analysis And Modeling Of The Atmospheric Carbon Dioxide In The Middle East And Comparisons With USA, EU And South Korea SCIREA Journal of Environment Background and Data," 2017.

[3] U. Al-mulali, "Factors affecting CO2 emission in the Middle East: A panel data analysis," *Energy*, vol. 44, no. 1, pp. 564–569, Aug. 2012.

[4] T. A. Boden, G. Marland, and R. J. Andres, "Global, Regional, and National Fossil-Fuel CO2 Emissions," 2011. [Online]. Available: https://cdiac.ess-dive.lbl.gov/trends/emis/overview_2010.html.

[5] P. P. Tans and T.J. Conway, "Monthly Atmospheric CO2 Mixing Ratios from the NOAA CMDL Carbon Cycle Cooperative Global Air Sampling Network, 1968-2002. In Trends: A Compendium of Data on Global Change.," *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A.*, 2005. [Online]. Available: https://cdiac.ess-dive.lbl.gov/trends/co2/cmdl-flask/cmdl-flask.html.

[6] S. Farhani and J. Ben Rejeb, "Energy Consumption, Economic Growth and CO 2 Emissions: Evidence from Panel Data for MENA Region," *Int. J. Energy Econ. Policy*, vol. 2, no. 2, pp. 71–81, 2012.

[7] R. D. Wooten and C. P. Tsokos, "Parametric Analysis of Carbon Dioxide in the Atmosphere," *J. Appl. Sci.*, vol. 10, no. 6, pp. 440–450, Jun. 2010.

[8] M. A. Stephens, “EDF statistics for goodness of fit and some comparisons,” *J. Am. Stat. Assoc.*, 1974.

[9] T. W. Anderson and D. A. Darling, “Asymptotic Theory of Certain ‘Goodness of Fit’ Criteria Based on Stochastic Processes,” *Ann. Math. Stat.*, 1952.

[10] H. Chernoff and E. L. Lehmann, “The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit,” *Ann. Math. Stat.*, 1954.

[11] D. D. Dunlop and A. C. Tamhane, *Statistics and Data Analysis: From Elementary to Intermediate*. 2001.

[12] U. Grömping, “Relative Importance for Linear Regression in *R* : The Package **relaimpo**,” *J. Stat. Softw.*, 2006.

[13] Y. Xu and C. P. Tsokos, “Attributable Variables with Interactions that Contribute to Carbon Dioxide in the Atmoshpere,” *Front. Sci.*, vol. 3, no. 1, pp. 6–13, 2013.

[14] I. Teodorescu and C. Tsokos, “Contributors of carbon dioxide in the atmosphere in Europe: the surface response analysis,” pp. 1–9.

[15] D. Kim and C. P. Tsokos, “Statistical significance of fossil fuels contributing to atmospheric carbon dioxide in South Korea and comparisons with USA and EU,” *J. Appl. Stat. Sci.*, 2013.

[16] M. I. Habadi and C. P. Tsokos, “Statistical Forecasting Models of Atmospheric Carbon Dioxide and Temperature in the Middle East,” *J. Geosci. Environ. Prot.*, vol. 05, no. 10, pp. 11–21, Oct. 2017.

[17] T. A. Boden, G. Marland, and R. J. Andres, “Global, Regional, and National Fossil-Fuel CO2 Emissions,” *The World Bank*. 2016.

[18] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting & Control*. 2015.

[19] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. 2013.

[20] S. H. Shih and C. P. Tsokos, “Prediction Models for Carbon Dioxide Emissions and the Atmosphere.”

[21] H. Akaike, “A New Look at the Statistical Model Identification,” *IEEE Trans. Automat. Contr.*, 1974.

[22] S. H. Shih and C. P. Tsokos, “A Temperature Forecasting Model for the Continental United States.”

[23] C. P. Tsokos, “Mathematical and Statistical Modeling of Global Warming,” in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 781–786.

[24] “2018 ALZHEIMER’S DISEASE FACTS AND FIGURES Includes a Special Report on the Financial and Personal Benefits of Early Diagnosis.”

[25] “Alzheimer’s Statistics – United States & Worldwide Stats.” [Online]. Available: https://braintest.com/alzheimers-statistics-throughout-the-united-states-and-worldwide/. [Accessed: 14-Mar-2019].

[26] C. Davatzikos and X. Da, “SPARE-MCI Scores from UPENN / SBIA : MRI-based biomarker of conversion from MCI to AD,” pp. 3–5, 2013.

[27] C. Davatzikos, F. Xu, Y. An, Y. Fan, and S. M. Resnick, “Longitudinal progression of Alzheimer’s-like patterns of atrophy in normal older adults: the SPARE-AD index.,” *Brain*, vol. 132, no. Pt 8, pp. 2026–35, Aug. 2009.

[28] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski, “Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification,” 2011.

[29] R. M. Chapman *et al.*, “Women have farther to fall: Gender differences between normal elderly and Alzheimer’s disease in verbal memory engender better detection of Alzheimer’s disease in women,” *J. Int. Neuropsychol. Soc.*, 2011.

[30] D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, and U. S. Army Academy, *Applied Logistic Regression Third Edition*. 2013.

[31] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. .

[32] A. AGRESTI, *introduction_to_categorical_data_analysis_805.pdf*, SECOND. .

[33] Z. Maintainer and D. Zhang, "Package 'rsq' Title R-Squared and Related Measures," 2018.

[34] D. Thompson and A. Health, "Paper D10-2009 Ranking Predictors in Logistic Regression."

[35] I. P. Bhatti, H. D. Lohano, Z. A. Pirzado, and I. A. Jafri, "A Logistic Regression Analysis of the Ischemic Heart Disease Risk," *J. Appl. Sci.*, vol. 6, no. 4, pp. 785–788, Apr. 2006.

[36] M. Habadi and C. P. Tsokos, "Alzheimer ' s : Probabilistic Approach to the Behavior of Beta-Amyloid and Phosphorylated Tau Proteins Levels," pp. 1–23.

[37] G. S. Bloom, "Amyloid-β and Tau: the trigger and bullet in Alzheimer's disease," *JAMA Neurol.*, 2014.

[38] P. A. Thomann, E. Kaiser, P. Schönknecht, J. Pantel, M. Essig, and J. Schröder, "Association of total tau and phosphorylated tau 181 protein levels in cerebrospinal fluid with cerebral atrophy in mild cognitive impairment and Alzheimer disease," *J. Psychiatry Neurosci.*, 2009.

[39] S. Mondragón-Rodríguez, G. Perry, X. Zhu, and J. Boehm, "Amyloid beta and tau proteins as therapeutic targets for Alzheimer's disease treatment: Rethinking the current strategy," *International Journal of Alzheimer's Disease*. 2012.

[40] M. S. Mendiondo, J. W. Ashford, R. J. Kryscio, and F. A. Schmitt, "Modelling mini mental state examination changes in Alzheimer's disease," in *Statistics in Medicine*, 2000.

[41] H. Joe, *9781466581432_googlepreview.pdf*. 1997.

[42] B. Rayens and R. B. Nelsen, "An Introduction to Copulas," *Technometrics*, 2000.

[43] K. As, "Modelling the dependence structure of financial assets: A survey of four copulas," *Samba*, 2004.

[44] U. Schepsmeier and E. Brechmann, "CDVine: Statistical inference of C-and D-vine copulas," *R Packag. version*, vol. 52, no. 3, 2012.

[45] K. Goda and S. Tesfamariam, "Multi-variate seismic demand modelling using copulas: Application to non-ductile reinforced concrete frame in Victoria, Canada," *Struct. Saf.*, 2015.

[46] M. Mahfoud and M. Massmann, “Bivariate Archimedean copulas: an application to two stock market indices,” *Vrije Univ. Amsterdam BMI*, 2012.

[47] I. Kojadinovic and J. Yan, “Modeling multivariate distributions with continuous margins using the copula R package.,” *J. Stat. Softw.*, vol. 34, no. 9, pp. 1–20, 2010.

[48] J. Yan, “Enjoy the Joy of Copulas: With a Package **copula**,” *J. Stat. Softw.*, vol. 21, no. 4, 2007.

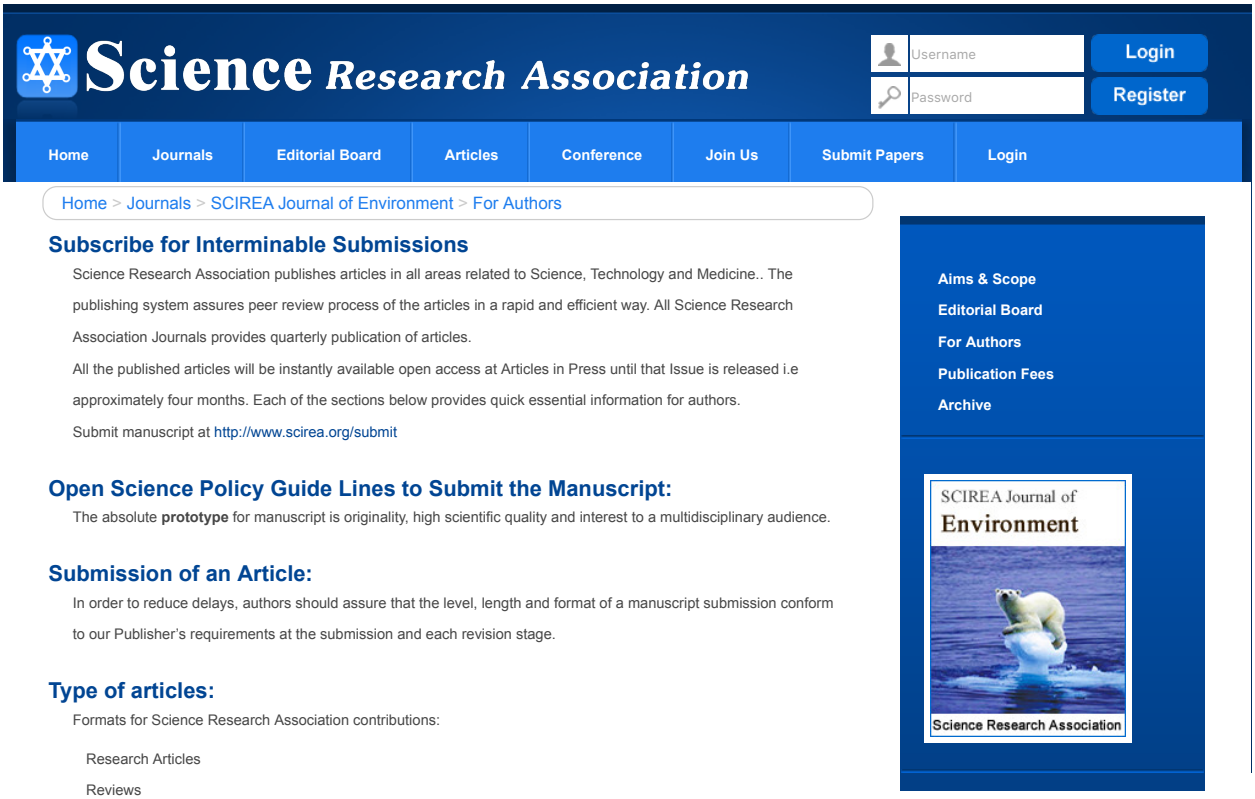
[49] T. T. Takeuchi, “Constructing a bivariate distribution function with given marginals and correlation: Application to the galaxy luminosity function,” *Mon. Not. R. Astron. Soc.*, 2010.

[50] X. S. Tang, D. Q. Li, C. B. Zhou, and L. M. Zhang, “Bivariate distribution models using copulas for reliability analysis,” *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, 2013.

[51] I. Wolfram Research, “Mathematica Online,” *Champaign, Illinois*, 2018. [Online]. Available: https://www.wolfram.com/mathematica/.

Appendices

Appendix A: Permission of Chapter Two



Science Research Association

Username Password Login Register

Home | Journals | Editorial Board | Articles | Conference | Join Us | Submit Papers | Login

Home > Journals > SCIREA Journal of Environment > For Authors

Subscribe for Interminable Submissions

Science Research Association publishes articles in all areas related to Science, Technology and Medicine.. The publishing system assures peer review process of the articles in a rapid and efficient way. All Science Research Association Journals provides quarterly publication of articles.

All the published articles will be instantly available open access at Articles in Press until that Issue is released i.e approximately four months. Each of the sections below provides quick essential information for authors.

Submit manuscript at http://www.scirea.org/submit

Aims & Scope
Editorial Board
For Authors
Publication Fees
Archive

SCIREA Journal of Environment

Science Research Association

Open Science Policy Guide Lines to Submit the Manuscript:

The absolute **prototype** for manuscript is originality, high scientific quality and interest to a multidisciplinary audience.

Submission of an Article:

In order to reduce delays, authors should assure that the level, length and format of a manuscript submission conform to our Publisher's requirements at the submission and each revision stage.

Type of articles:

Formats for Science Research Association contributions:

- Research Articles
- Reviews
- Abstracts
- Book Reviews
- Rapid Communications
- Letters to the Editor
- Case Reports
- Meeting Reports
- Orations
- Product Reviews
- Founders' Reviews
- Breakthrough Technologies
- Hypotheses and Analyses

Author Guidelines:

All manuscripts must be submitted through the journal's online submission at http://www.scirea.org/submit

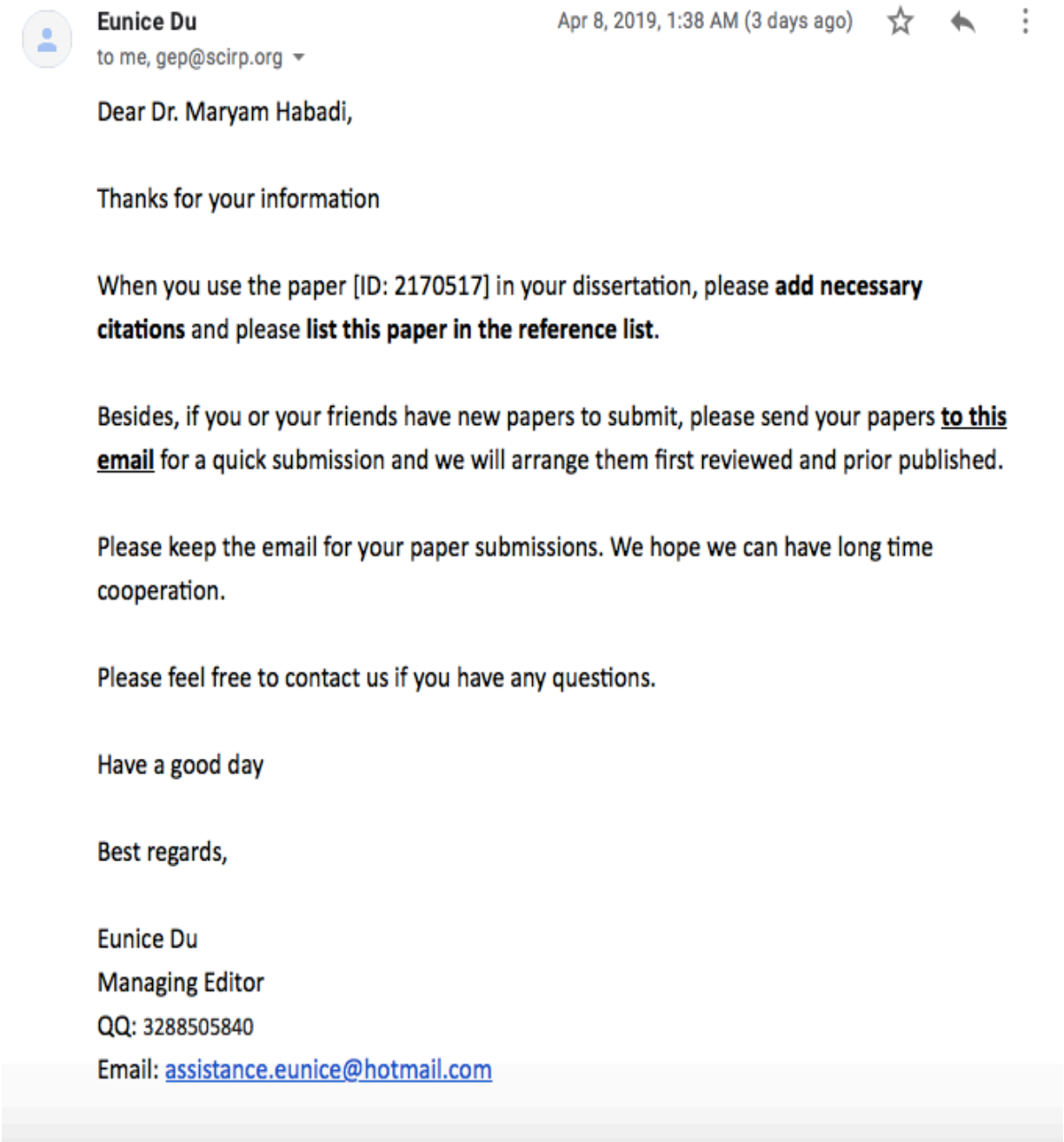
Our aim is to provide all authors with an efficient, courteous, and constructive editorial process. Contributions should therefore be written clearly and simply so that they are accessible to readers in other disciplines and to readers for whom English is not their first language. Essential but specialized terms should be explained concisely but not didactically.

Criterion for the Article:

We will consider manuscripts of any length; we encourage the submission of both substantial full-length bodies of work and shorter manuscripts that report novel findings that might be based on a more limited range of experiments. The key criteria are that the work clearly demonstrates its novelty, its importance to a particular field as well as its interest to those outside that discipline, and conclusions that are justified by the study.

Copyright

Appendix B: Permission of Chapter Three



Eunice Du Apr 8, 2019, 1:38 AM (3 days ago)

to me, gep@scirp.org

Dear Dr. Maryam Habadi,

Thanks for your information

When you use the paper [ID: 2170517] in your dissertation, please **add necessary citations** and please **list this paper in the reference list**.

Besides, if you or your friends have new papers to submit, please send your papers **to this email** for a quick submission and we will arrange them first reviewed and prior published.

Please keep the email for your paper submissions. We hope we can have long time cooperation.

Please feel free to contact us if you have any questions.

Have a good day

Best regards,

Eunice Du
Managing Editor
QQ: 3288505840
Email: assistance.eunice@hotmail.com